Document description languages

SGML,HTML,XML

SGML
(Standard General Markup Language)

- A great variety of documents (ex. web): articles, catalogoues, lists, data tables etc..

- Each document has its logical structure (article: title , author, data,..) . The standard format used for all documents is ASCII, but different conventions are used in representing information (ex.,name and surname or viceversa, different number of bytes reserved , name delimited by "$"..).

- The management of the archive is difficult. (searching for a text, for the author name or for all its the books,..)

- It is necessary to adopt a standard markup language. A markup language defines how documents shoud be formatted.

•In the early days of computer type setting, there were many different typsetting systems, and each used its own proprietary markup language. This language consisted of special control characters to indicate the beginning and the end of some formatting.

•**SGML** (**S**tandard **G**eneral **M**arkup **L**anguage).
developped in 1986 by International Organization for Standardization (ISO)
.

•**SGML is a metalanguage**: a language that describes a formatting and markup language.

•**HTML** (**H**yper **T**ext **M**arkup **L**anguage) is the first simplified language deriving by SGML It is characterize by a set of **tags** and rules for their use.

HTML(Hypertext Markup Language)

- HTML is a *markup* language because is not include detailed formatting information.
- For example, although HTML contains extensions that allow an author to specify the size of the text, the font to be used or the width of a line, most authors choose instead to specify only a level of importance as a number from 1 to 6.
- The browser choses a font and disply size appropriate for each level.
- Simarly HTML does not specify exactly how a browser marks an item as selectable. Some browsers underline selectable items, others display selectable items in a different color and some do both.
- Consequently two browsers may display an HTML document differently

- Syntactically, each HTML document is represented as a text file that contains **tags.**

- **HTML tags provide structure for the document as well as formatting hints.**

- Some tags specify an action that takes effect immediately (e.g., move to a new line on the display); the tag isplaced exactly where the action should occur.

- Other tags are used to specify a formatting operation that applies to all text following the tag. Such tags occur in pairs, with a leading tag and a trialing thread tha start and terminate the action, respectively .

•A tag appears as a **tag name** bracketed by *less-than* and *greater than* symbols:

<TAGNAME>

•The corresponding tag used to end an operation begins with two-character sequence *less-tha*n and *slash* and ends with a *greater than-symbol:*

</TAGNAME>

- General form of a HTML document

```
<HTML>
        <HEAD>
                <TITLE>
                text that forms the document title
                </TITLE>
        </HEAD>
        <BODY>
                body of the document appears here
        </ BODY >
</HTML>
```

Note that an indentation is used to show the structure. However , the browser ignores all such spacing.

Examples HTML formatting tags

Hello there. <BR>This is an example<BR>of HTML

Hello there.
This is an example
of HTML

Hello there. <BR><BR>This shows <BR> HTML spacing

Hello there.

This shows
HTML spacing

Headings

- Html contains **six pairs** of tags that can be used to display heding in the output.

A tag of the form **<Hi>** marks the start of a level i heading, and a tag of the form **</Hi>** marks the end.

- Text with te most important level of heading is bracketed between <H1> and </H1>.

- Example:

```
Hello.<BR><H1> This is A Heading </H1><BR> Back to normal
```

Hello

This Is A Heading

Back to normal

LISTS

HTML allows a document to contain lists. The simplest form is an *unordered list*, which requests the browser to display a list of items.

```
Here  is a list of 5 names:
<UL>
<L1> Scott
<L1> Sharon
<L1> Jan
<L1> Stacey
<L1> Rebecca
</UL>
This text occurs after the list
```

Here is a list of 5 names

- Scott
- Sharon
- Jan
- Stacey
- Rebecca

This text occurs after the list

```
<HTML>
                    <HEAD>
                    <TITLE> First example </TITLE>
                    </ HEAD>
                    <BODY>
                     First example of a web page written in XML . Even if the text
                    begins a new paragrafh, to begin a new paragraph in the
                    displyed page it is necessary to use the tag<BR>oppure il tag
                    <P> that also to begin a new paragraph  inserts an empty row
                    </BODY>
</HTML>
```

First example of a web page written in XML . Even if the text begins a new paragrafh, to begin a new paragraph in the displyed page it is necessary to use the tag
or the tag

that also to begin a new paragraph inserts an empty row.

tags

<CENTER>…</CENTER> the content is inserted in the center of the window

<OL>…</OL>.ordered lists

<HR WIDTH="100%"> an horizontal line large as the window is created

```
<HTML>
<HEAD><TITLE> examples </TITLE></HEAD>
 <BODY BGCOLOR=Yellow>
<H1><CENTER> centered title </CENTER> </H1>
<H2><CENTER> centered sub-title </CENTER> </H2>
<H3><CENTER> centered sub-sub-title </CENTER> </H3>
<HR WIDTH= “100”%>
<H2> unordered list </H2>
<UL>
<LI> First element </LI>
<LI> Second element </LI>
<LI> Third element </LI>
</UL>
<HR WIDTH= “100”%>
<H2> ordered list </H2>
<OL>
<LI> First element </LI>
<LI> Second element </LI>
<LI> Third element </LI>
</OL>
</BODY>
</HTML>
```

Centered title

Centered sub-title

Centered sub-sub-title

----------------- --

not ordered list

- First element
- Second element
- third element

----------------- ---

ordered list

1. Firts element
2. Second element
3. third element

Embedding Graphics Images in A Web Page

- Non textual information such as **graphical information or a digitized photo** is not inserted directly in a HTML document. Instead the data resides in a separate location, and the document contains a reference to the data.

- When a browser encounters such a reference, the brower goes to the specified location, obtains a copy of the image and inserts the image in the displayed document.

```
<IMG SRC= “fred_photo.gif”>
```

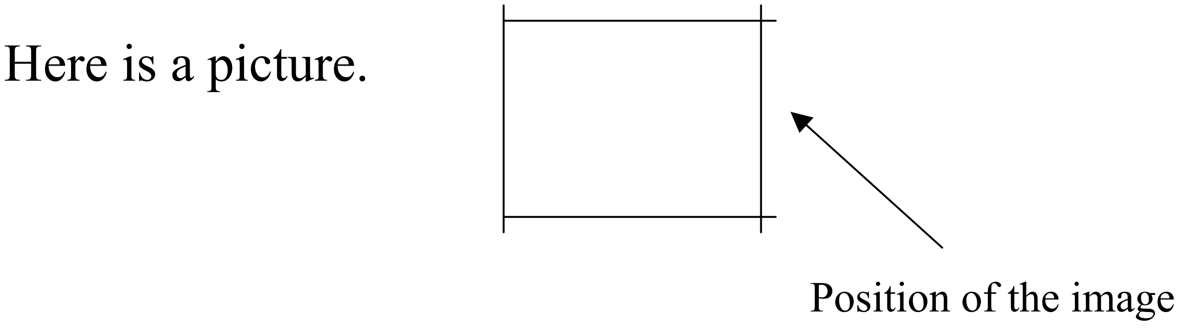
specifies that the file *fred_photo.gif* contains an image that the browser should insert in the document.

Image files are not stored as text files. Each image file contain binary data and the file is stored in *graphics interchange format (gif)*

- IMG tag includes additional parameters that can be used to suggest positioning.

- The keyword ALIGN can be used to specify whether the top, middle, or bottom of the image should be aligned with others items on the line.

Here is a picture. <IMG SRC=“ fred_photo.gif” ALIGN=middle>.



Hypertext Links from one document to Another

The HTML mechanism for specifyng a hypertext reference is known as an *anchor.*

To permit arbitrary text and graphics to be included in as inglereference, HTML uses tags <A> and </A> as a bracket for the reference.

For example:

```
This Book is published by
<A HREF=://www.prenhall.com”>
Prentice Hall,</A> one of
the larger publishers of Computer Science textbooks.
```

When displayed the input produces:

This book is published by Prentice Hall ,one of the larger publishers of Computer Science textbooks.

HTML limitations

- The tags of the language are fixed and no modifiable . Also,it is oriented to the description of hypertext documents.

- Partially structured. It depends from the characteristics of the used browser.

- A web page must be designed for a particular display characterized by specific features.

•As the web grew in popularity, HTML was extended for new purposes; however , soon it becames apparent that proprietary extensions to HTML were counter-productive and ill-suited to general use.

• For each new version of browser a propritary extension of HTML was proposed. As a result of such situation, it was impossible to display the same pages of a web site on different browsers

• So, to solve the problems of interoperability and scalabilty on the web without extending HTML, the W3C began work on a simplified version of SGML which is called the **Extensible Mark-up Language (XML).**

However, even if the initial objectives of XML were relative to the solution to a standard problem for the web, XML is not limited only to web context.

It is used primarily for the exchange of data between two network applications. It allows , moreover , the representation of complex data types.

• A XML document is a text file which contains tags, attributes and text following well defined syntactic rules .

• XML logical structure
A XML document has a hierarchical structure. Its components are called elements. Each element represents a logical component of the document and may contain others elements of the text.

•Informations can be associated to the elements, describing their properties. These informations, in the form of a name/value, are called attributes..

• The elements are hierarchically organized with a principal element called root element.

• The root contains the set of others elements of the document l'insieme degli altri elementi del documento. From a graphical point of view thestrucure can be represented as a treee (document tree).

Some syntactic differences between XML and HTML

• XML elements (composed of a start and end tag) must be stricly nested.
<B><I> improper nesting </B></I> is legali i n HTML, but illegal in XML.

• In XML every start tag must have an end tag.

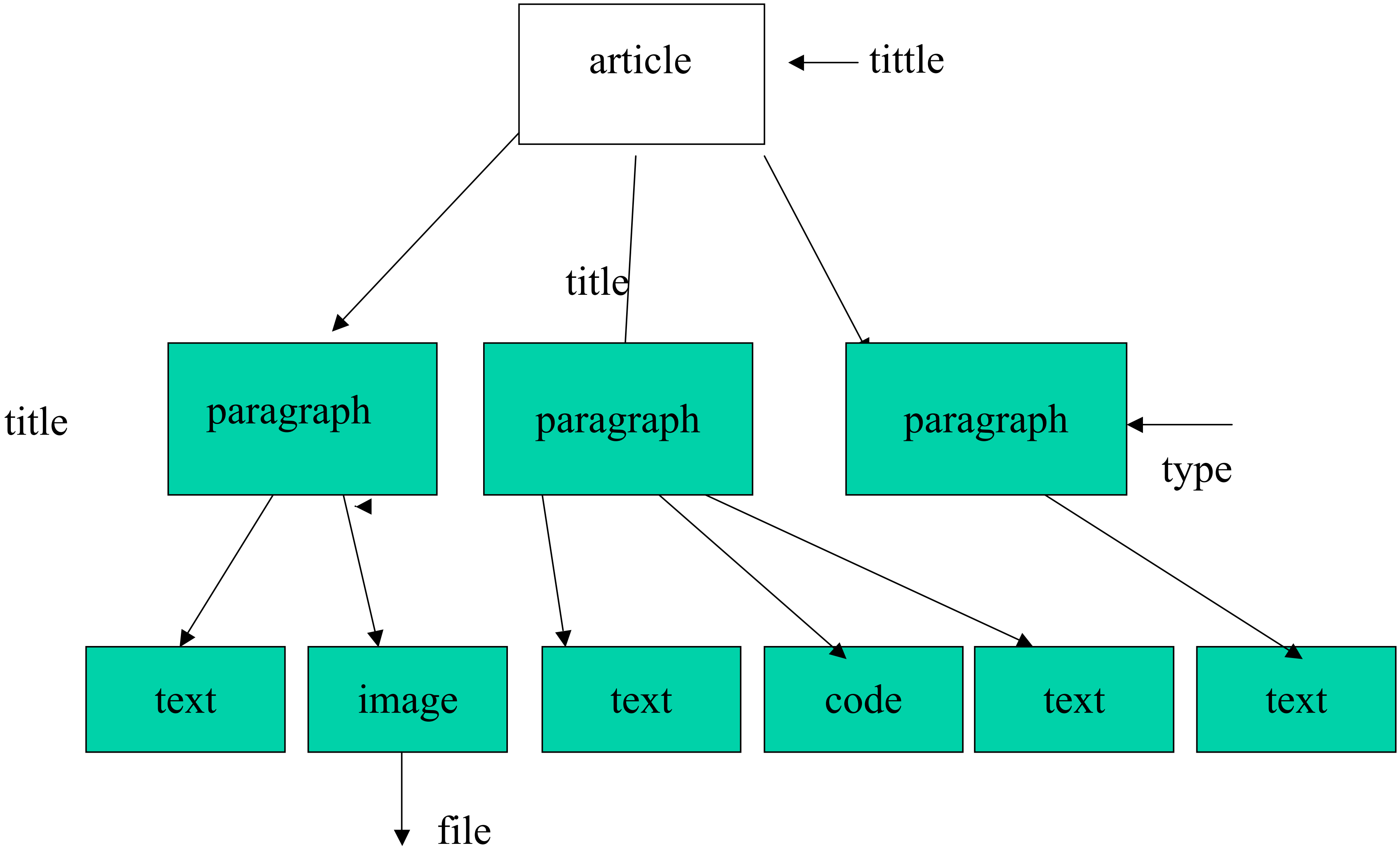
• XML documents allow only one root element, the top level element that contains all other elements.

• All attributes values in XML must be sorrounded by single or double quotes.

• XML tags are case sensitives.

• Whitespace between tags is ignored in HTML, but is preserved in XML and considered relevant.

•**Well formed document.** A XML document is well formed if it follows all the syntax rules scontained in the XML specification


article
tittle
title
title
paragraph
paragraph
paragraph
type
text
image
text
code
text
text
file

- In the figure we have a root element, article, which contains a list of elements representing the paragraphs of the article. Each paragraph contains texts, programs and images. Some elements are characterized by attributes: title,type, file

- The logical structure of a XML document depends from the type of organization of the elements. There are not univeral rules to decide the logical organization of a document. It depends by our decisions.

- The logical structure of a XML document is translated in a corresponding phisical structure constituited of syntactic elements called **tags.**

- The phisical representation of the XML document is the following:

```
<?xml version="1.0" ?>
<article title=“Article title">
  <paragraph title=“title of the first paragraph ">
    <text>
      text block of the first paragraph
   </text>
    <image file="image.jpg">
    </image>
  </ paragraph >
  < paragraph titolo="Title of the second paragraph”>
    <text>
      text block of the second paragrah
    </text>
    <code>
      program
    </code>
    <text>
      Anther text block
    </text>
  </paragraph>
  <paragraphe type="bibliography">
    <text>
      reference to an article
    </text>
  </paragraph>
</article>
```

```
<?xml version="1.0" encoding = ISO-8859-1">
<music>
  <producer>
     <name> Karim</name>
     <city> Roma</ city>
     <catalog>
       <disc year= "1961">
          <title>Nuvole Barocche </title>
          <singer>Fabrizio De Andrè </singer>
       </disc>
       <disc year= "1965">
          <title> La città vecchia </title>
          < singer >Fabrizio De Andrè </ singer >
       </disc>
         ……
    </catalog>
 </producer>
        ……..
</music>
```

•The elements and the attributes represent the key indicators of the structure or purpose of our content. We must now to determine which tags we can use in a document.

• In other words we must define a **grammar** for the particular markup language
A grammar is a set of rules that defines the.words (elements) and the structure by which it is possible to construct sentences (documents).

• A grammar defines a specific markup language . If a XML document respects the rules defined by a grammar it is **valid** for the corresponding language.

• To be automatically elaborated a XML document must be well formed and valid.

Components of XML

- XML is a formalization of rules for “marking up” documents. There are six types of **markup:elements, attributes, comments, processing instructions, entity references and CDATA sections**.

- **Elements**. Are the more common aspect of markup languages. An element is a logical construct of a document. A normal element is composed of a start and end tags that surround content, others elements or both..

```
<street> 4296 Razor Hill Road</street>
```

Attributes . An element may have attributes that are specified in name/value pairs and are placed after the start-tag name. In this example the width and height are the attributes:

<Applet width=“100” height=“200”>

Comments

A comment allows fre text description tha it is ignored by an XML processor. For example:

<!– Keep this part is really important.-->

Processing instructions

Are used to pass information to a processing application. Example:

<?application data?>

Entity references.

Entity references are used to put reserved characters or abbreviations in markup. For example, the left angle bracket(<) is a reserved character.

CDATA Sections.

A Cdata section is a section of text that shoud not be processed but instead passed directly to theapplication. Thi is useful for passing source code to an applicationpass

DTD (Document Type Definition)

- Tag names are not fixed in XML. Only syntactic rules are defined relatives to their definition and use.

- A DTD declares all the legal elements in a document; the legal attributes those elements can have; and the hierarchy, nesting and occurence indicators for all elements. In order for a document to be valid it must specify what DTD it adheres to.

- A XML document that satisfies the specifications of a DTD is validated with reference to that DTD.

- Automatic tools exist to validate a XML document.

The class of documents that are consistent with the document structure may be defined by the following DTD :

```
<!ELEMENT music (producer +)>-
<!ELEMENT producer (name, city, catalog)>
<!ELEMENT name (#PCDATA)
<!ELEMENT city (#PCDATA)
<!ELEMENT catalog (disc+)>
<!ELEMENT disc (title,singer)>
<!ATTLIST disc year CDATA#REQUIRED>
<!ELEMENT title (#PCDATA)>
<!ELEMENT singer (#PCDATA)>
```

Music is the root element. It contains a list of producers (one or more).
Each producer is constituted by three elements: :*name, city and catalog*.
The first two (as in the following, *title and singer*) contain only text .(PCDATA), instead *catalog* contains a list of one or more disc elements, each of them with the attribute year (REQUIRED, cannot be omitted) , a title and the singer name.

EXAMPLE

```
<Phone _book>
   <item>
       <Name> Mario Bianchi </Name>
       <Number> 0665745689 </Number>
       <Address> street della viola 37  00132 Roma </Address>
  </item>
…………..
 <item>
     <Name> Sandro Verdi </Name>
     <Number>0235769856</Number>
     <Address>via delle rose 63  20127 Milano </Address>
</item>
</Phone_book>
```

DTD

```
<!DOCTYPE phonebook
          <!ELEMENT phonebook (VOICE+)>
           <!ELEMENT item (NAME, NUMBER, ADDRESS)>
          <!ELEMENT NAME (#PCDATA)>
          <!ELEMENT NUMBER (#PCDATA)>
          <!ELEMENT ADDRESS(#PCDATA)>
          ]>
```

Standard for the definition of XML application domains

•In order to assigne a semantic value to the elements and attributes of a XML document , standards have been created for different application domains.

SBML (System Biology Markup Language)(http://sbml.org)
GML (Geography Markup Language)(http://www.opengeospatial.org)
HealthCareLevel Seven (http://www.hl7.org)
XBRL (XML based Business Reporting standard (http://www.xbrl.org)
GJXDM (Global Justice XML Data Model) (http://it.oip.gov/jxdm).

Rosetta net consortium has defined a number of document types and their semantic to use in B2B transactions in the ICT sector.

Al momento è necessario (web services) concordino sull’uso dei termini della transazione che si vuole realizzare.

XML offre solo interoperabilità di tipo sintattico e strutturale, ma non una reale condivisione di conoscenza, quando non vi sia già una semantica condivisa.

Ontologie.

Web semantico

Query and transformations languages

Tranformation

•The browser cannot interprete the tags (differently from HTML).

• XML does not allow the description of the graphic presentation of the logical elements of the text. Special purpose languages (stylesheets) are used.

•It is possible to obtain from the same XML document different kinds of pubblication (paper www, audio,..) by using different style*sheets*

• A common stylesheets is *XSL (eXtensible Stylesheet Language)*. It is possible not only to decide the graphic format, but also to decide which parts of the document must be displayed.

Query

Xpath.
Sintax like to file pathname in order to find element depending on their position in the XML tree.
Ex::

```
doc (music.xml)//catalog/disc
```

Estracts all the disc elements contained in the catalog element of the xml music document

```
doc (music xml)//catalog/disc [singer “Bob Dylan”]

doc (music.xml)//catalog/disc[singer“Bob Dylan”]/title
```

Parsing

•Parsing is the process of dissecting a body of text into its individual component pieces.

• For example, if that body of text is a paragraph, parsing will break the paragraph into sentences. It would then break a sentence into subject and predicate. In turn, the subject and the predicate would then be broken down into their components like nouns, verbs and adjectives.

•Lexical analysis breaks the body of text into tokens. Tokens are the smallest atomic components of the stream of data. In the paragraph example, tokens would be words (scanner)

•Grammatical analysis. Involves recognizing the syntactical structure of a language. In other words , how words are combined to form larger structures and howthose structures form even larger ones (parser).

•SAX, DOM

Base Software

- XML Browser. Allows the simple reading of XML documents Open source, Internet Explorer..,

- Validationion parser and XSLT are available as open source products

- XML Editor (file di testo) : provide features for syntactic validation and DTD validation

- MS XML Notepad, XML Pro,ect. XML SPY

XMLand data-base

- It is possible to obtain a XML form of a query possibile esporre in formato XML il risultato delle interrogazioni. La codifica di una tabella relazionale in XML è semplice.

- Example: Microsoft SQL Server allows to obtain the XML form of a query by using “FOR XML”

- It is possible to modify a relational database using XML data..

XHTML

(e**X**tensible **H**ypertext **M**arkup **L**anguage)

• Sfruttando le somiglianze sintattiche è stato definito un linguaggio di markup per le *pagine web* che mette a disposizione le possibilità di *HTML* con una sintassi *XML*

•Semplifica la programmazione di browser per computer e cellulariSoftware di base

Analogie e differenze con HTML

•Entrambi derivano da SGML, ma HTML è stato specializzato nella parte relativa alla visualizzazione dei dati.

•XML non è orientato alla visualizzazione dei dati, ma è un metodo assolutamente generale di descrivere i dati.

•Somiglianze sintattiche (i tag sono indicati allo stesso modo, <nome tag> </nome tag>).

•A differenza di HTML, in XML i tag sono "liberi", cioè definiti dal programmatore.

•Sintassi di XML più rigida (ogni tag aperto deve essere anche chiuso, vi è differenza tra maiuscole e minuscole..).

•XML ha una struttura ad albero , cioè gerarchica.

DTD (Document Type Definition)

•XML *non prescrive* i nomi dei diversi marcatori, ma solo la *sintassi generica* per la loro definizione ed il loro utilizzo nella identificazione degli elementi di testo.

• DTD è un *insieme di specifiche* che stabiliscono quali sono i *nomi ammissibili per i marcatori,* i nomi per i loro *attributi* e quali relazioni di *inclusione* possono sussistere tra loro.
Descrive rigorosamente le possibilità strutturali del documento in esame.

•*Un documento XML che* soddisfa le specifiche di una DTD si dice *validated* (convalidato) rispetto a quella DTD.

Esistono degli strumenti automatici che possono verificare se un documento è consistente con quanto è prescritto dalla sua DTD.

•La DTD può essere inserita all’inizio del documento XML o memorizzata in un file diverso cui fa riferimento.

•La DTD può essere privata (scritta dall’utente stesso) oppure pubblica (reperita in rete).

•La DTD di un documento XML **non fornisce alcuna informazione semantica**: definire che un certo elemento deve essere contenuto in un documento XML *non è sufficiente per chiarire la sua semantica e come l’informazione verrà utilizzata da chi riceve il documento*.

•Quando l’informazione viene estratta, deve esserci un programma che comprende la semantica.

Ontologie

Specifiche formali di concettualizzazioni che descrivono una comprensione comune di un dominio, la quale è concordata da una pluralità di soggetti e può essere deliberatamente condivisa tra persone diverse ed applicazioni diverse.

Semantic web

Affidare al meccanismo di scambio di dati alla base dei web services anche una descrizione dei domini realizzata tramite ontologie.

Standard per la descrizione di domini applicativi in XML

•E’ possibile,da parte di consorzi di standardizzazione, definire tipi di documenti XML (tramite DTD e schemas) in modo che possa essere fissata la semantica di ciascuno degli elementi e attributi di un documento che viene scambiato tra le parti interessate.

SBML (System Biology Markup Language)(http://sbml.org)
GML (Geography Markup Language)(http://www.opengeospatial.org)
HealthCareLevel Seven (http://www.hl7.org)
XBRL (XML based Business Reporting standard (http://www.xbrl.org)
GJXDM (Global Justice XML Data Model) (http://it.oip.gov/jxdm).

Rosetta net consortium ha definito un numero di tipi di documenti e la loro semantica per transazioni B2B nel settore IT.