XML: Current Developments and Future Challenges for the Database Community

Stefano Ceri, Piero Fraternali, Stefano Paraboschi

Thoughts….

- Comparison to Ashish
- EXtending database technology
- Picking up Peter's recommendation

XML

http://w3c.org/XML/

- eXtended Markup Language
- Document markup language for the World Wide Web proposed by the W3C
- Examples of electronic documents: books, manuals, product catalogs, messages, news, math formulas…

Evolution

- 1986: Standard Generalized Markup Language (SGML) ISO 8879-1986
- November 1995: HTML 2.0
- January 1997: HTML 3.2
- August 1997: XML Working Draft
- December 1997: XML 1.0 Proposed Recommendation

XML vs HTML

- HTML: fixed tags, mostly concerned with text presentation
- XML: domain-specific tags, concerned only with document semantics

```
<h1> EDBT 2000 </h1>
 <ul>
  <li> C. Zaniolo P. C. Lockemann
  <li> M.H. Scholl, T. Grust (eds.)
  <li> Springer-Verlag
 </ul>
```

```
<book>
<title>EDBT 2000 </title>
 <editor> C. Zaniolo </editor>
 <editor> P. C. Lockemann </editor>
 <editor> M. H. Scholl </editor>
 <editor> T. Grust </editor>
 <publisher> Springer-Verlag
    </publisher>
</book>
```

Why XML is good news?

- Data semantics - XML documents are self-describing and must comply with a document type definition (DTD) that dictates the schema of the document.
- Data independence - XML documents are specified independently from presentation.

The consequence

- **XML turning from a pure document markup language into a data interchange format**
  - **An instrument for enabling data publication by various applications that need to co-operate.**
  - **Key enabling concept for achieving data interoperability.**

http://w3c.org/XML

- Enable internationalized media-independent electronic publishing.
- Allow industries to define platform-independent protocols for the exchange of data, especially the data of electronic commerce.
- Deliver information to user agents in a form that allows automatic processing after receipt.
- Make it easy for people to process data using inexpensive software.
- Allow people to display information the way they want it.
- Provide metadata -- data about information -- that will help people find information and help information producers and consumers find each other.

Talk Outline

- XML as a data representation standard
- XML as a data interchange standard
- Repository technology for XML
- Research experiences with XML
  - Query language (XML-GL)
  - Active document management
  - Model-driven conceptual Web design (WebML)

First viewpoint: XML as a Data Representation Standard

- Classical data abstractions:
  - Modeling
  - Querying
  - Updating
  - Viewing and constraining
  - Mining

  …. apply to XML !

XML Data Modeling

- DTDs = grammar of XML data
  - with iteration, optionalities, alternatives
- DTDs = Object-Oriented schema (with containment)
  - Element = Object
  - ID Attribute = OID
  - IDREF, IDREFS Attributes = References to OID
  - Alternatives = Union type
- Missing features (the subject of XML Schema)
  - Class hierarchies
  - Types (and typed object references)
  - (some) Integrity constraints

DTD Design Problems

- Similar to schema design (for large collections of homogeneous documents)
  - Decide entities and relationships
  - Choose most relevant one-to-many relationships as containment hierarchies
  - Model the other relationships as IDREF links
- Conceptual difficulties
  - Using attributes vs elements for storing PCDATA content
  - PCDATA sub-structuring within a given element
  - Ordering
  - Integrity constraints (such as keys, referential integrity)
- Inferring the DTD of a given document

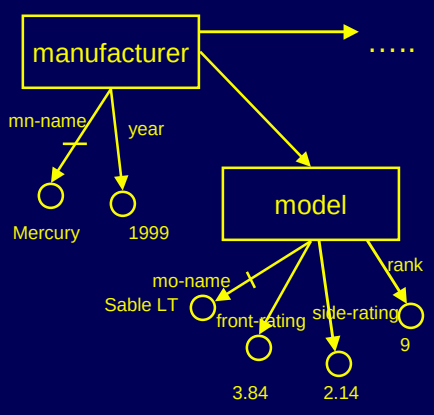
XML Query Languages

- The first XML query languages
  - LOREL (Stanford) - almost by accident
  - XSL, XQL
- W3C-related events
  - XML-QL (ATT-INRIA) as W3C request for standardization
  - W3C Workshop & Working Group - towards a QL standard (expected for November 2000)
  - W3C "QL requirements" document (out)
- Our (POLI) proposal
  - Graphical QL (XML-QL)

Comparative examples

```
<manufacturer>
 <mn-name>Mercury</mn-name>
 <year>1999</year>
 <model>
 <mo-name>Sable LT</mo-name>
 <front-rating>3.84</front-rating>
 <side-rating>2.14</side-rating>
 <rank>9</rank>
 </model>
 ……
</manufacturer>

<vehicle>
<vendor>Scott Thomason</vendor>
<make>Mercury</make>
<year>1999</year>
<model>Sable LT</model>
<color>metallic blue</color>
<option opt="sunroof">
 ……...
<price>26800</price>
</vehicle>
```



Lorel

- Developed at Stanford University (S. Abiteboul, J. McHugh, D. Quass, J. Widom, J. Wiener)
- User-friendly language in the SQL-OQL style, with:
  - very powerful path expressions
  - strong mechanism for type coercion

Example of query in Lorel

- Select and extract <manufacturer> elements where some <model> has <rank> less or equal to 10.

```
select M
from nhsc.manufacturer M
where M.model.rank<=10
```

XML-QL

- Designed at AT&T Labs (A. Deutsch, M. Fernandez,D. Florescu, A. Levy, D. Suciu):
- Imitates in the query the style of XML documents:
  - distinguishing features
    - makes an explicit **`construct`** clause to build the result
    - can express queries as well as transformations for integrating XML data from different data sources

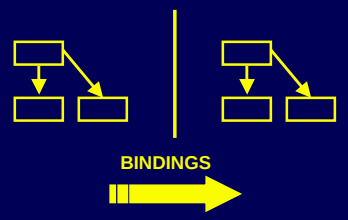
Example of query in XML-QL

- Select and extract <manufacturer> elements where some <model> has <rank> less or equal to 10.

```
WHERE <manufacturer>
        <model>
         <rank>$r</rank>
        </model>
       </manufacturer> ELEMENT_AS $m
          IN www.nhsc/manufacturers.xml,
       $r<=10
CONSTRUCT $m
```

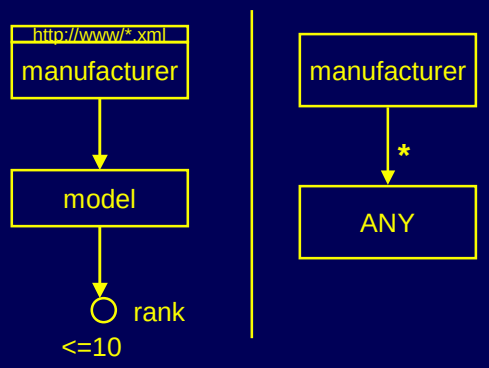
XML-GL

- Designed at Politecnico di Milano (S. Ceri, S. Comai, E. Damiani, P. Fraternali, S. Paraboschi, L. Tanca):
  - a graphical query language, relying on a graphical representation of XML documents by means of labeled *XML graphs*
  - suitable for supporting a user-friendly interface



Example of query in XML-GL

- Select and extract <manufacturer> elements where some <model> has <rank> less or equal to 10.



Extensible Stylesheet Language (XSL)

- Developed by the W3C XSL Working Group:
  - an XSL program is a collection of template rules; each template rule has a pattern which is matched against nodes in the source tree, and a template instantiated to form the result

Example of query in XSL

- Select and extract <manufacturer> elements where some <model> has <rank> less or equal to 10.

```
<xsl:template  match="/">
 <xsl:for each  select="manufacturer[model/rank<=10]">
     <xsl:value-of   />
 </xsl:for each>
</xsl:template>
```

XML Query Language (XQL)

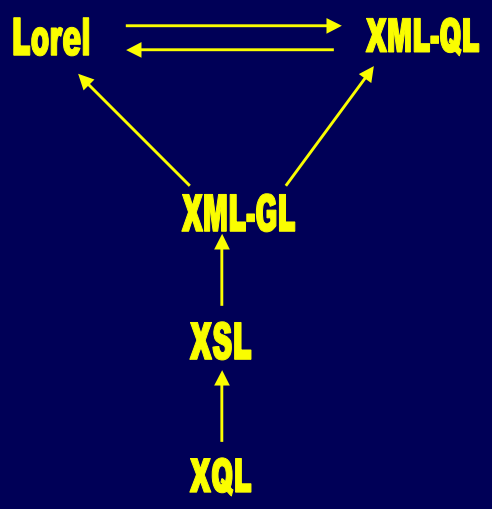
- Developed by J. Robie (Texcel Inc.), J. Lapp (webMethods Inc.) and D. Schach (Microsoft Corporation):
  - XQL expressions are easily parsed, easy to type, and can be used in a variety of software environments - as part of a URL, in XML or HTML attributes, in programming language strings, etc.

Example of query in XQL

- Select and extract <manufacturer> elements where some <model> has <rank> less or equal to 10.

```
manufacturer[model/rank<=10]
```

Classification



Comparison

- **Lorel** and **XML-QL** are the OQL-like and XML-like representatives of **Class 2** (SQL2) of expressive query languages for XML.
- **XSL** and **XQL** is representative of **Class 1** (core SQL) of single-document query languages.
- **XML-GL** enables a *graphical query interface*, and can be considered equivalent to a **QBE** for XML.

- For more info: ACM-SIGMOD RECORDS, March 2000.

Language-independent research

- **Define XML Algebra**
  - **Orthogonal & minimal algebraic operators**
  - **Define equivalence properties and high-level optimization**
- **Enhance query languages**
  - **Proximity search**
  - **Accept approximate results**
  - **Combine queries and keyword-based search**

Beyond QL

- Views (representing derived data)
  - A useful concept for building “derived sites”
  - Require materialization and incremental maintenance
- Semantic constraints
  - Referential integrity beyond containment
- Updates
  - High-level operations of insert-update-delete
- Triggers
  - Requires the notion of event and update

Second viewpoint: XML as a Data Interchange Standard

- Several successful data interchange standards
  - SQL (with JDBC): “intergalactic data speak”
  - CORBA, DCOM: “distributed components”
- But they do not solve many interoperability problems
  - They describe “computations” and not “data”
  - They don’t help in describing the semantics of data being exchanged between systems

XML-enabled data interchange protocols

- E-commerce protocols for negotiation and bidding
- Agent-based computations for automatic information discovery
- Improved (semantic) search engines
- Domain-specific semantic descriptions
  - Genetic, math, chemical data
  - XML-based computer systems specifications (XMI)

Beyond XML

- XML Schema - an extension of DTDs
  - Data types support
  - Generalization hierarchies
  - Typed links
  - Integrity constraints
- Very much complete (too much complete??)

Third viewpoint: XML as a Repository Technology

- Standard "Document Object Model" (DOM) Technology for storing and retrieval XML documents
  - Interfaced by means of standard XML parsers
  - First generation of XML servers based on DTD-independent files + indexing + text matching
  - Future versions could be more DTD-influenced and reuse services of object stores or relational storage servers

Challenges for "core DB" experts

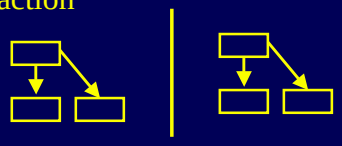
- Support XML queries by means of ad-hoc data structures & indexes
- Use DTDs knowledge to optimize queries
- Support parallel and distributed query processing
- Dealing with replicas and order
- Deal with irregular data and heterogeneous data sources

Research at Politecnico di Milano on XML

- **XML-GL**
- **ACTIVE XML-GL**
- **WEBML and W3I3**

XML-GL

- Already shown in action



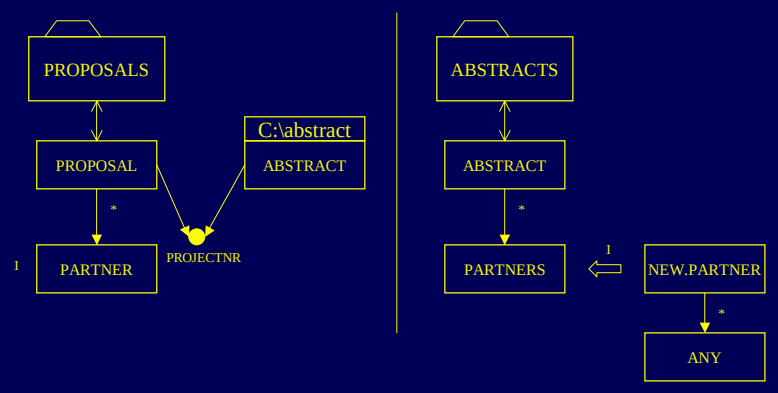
- Designed with the intention of becoming the QBE of QL for XML.
- Will adapt to the forthcoming W3C standard QL
- Various ongoing projects
  - Translation to LOREL
  - Translation to GMD-IPSI XQL Engine

Active rules for XML

- Active rules adopt the ECA paradigm
  - EVENT: a change on an XML element, possibly detected off-line
  - CONDITION: an XML query
  - ACTION: an update command on an XML document
- All results of previous research on active rules are applicable: termination analysis, confluence requirements, ...
- New problem: edit-script independence

Active XML-GL

- Whenever a new proposal is filed, extend the abstracts document with proposal's partners.



Rules in LOREL

- Lorel updates combined with Lorel queries give the condition-action part. It is sufficient adding:
  - Events
  - Binding passing mechanisms from events to actions

- An example:

```
Event: insert (proposal.#.partner PP)
Action: update A.#.partners.partner += PP
        from abstract A, proposal P
        where P.projectNr = A.projectNr and
              P.#.partner = PP
```

Applications of XML rules

- Several important applications
  - Document classification
  - Integrity maintenance
  - Push technology
  - Materialized view management
  - Derived data computation
  - Workflow management

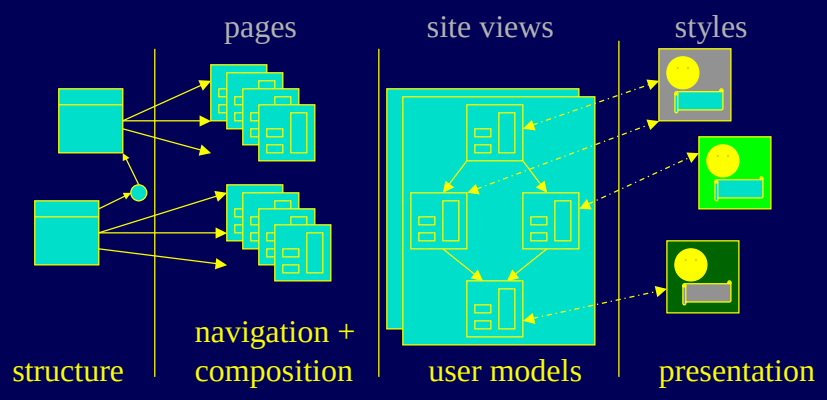
- Implementation currently ongoing (together with XML-GL)

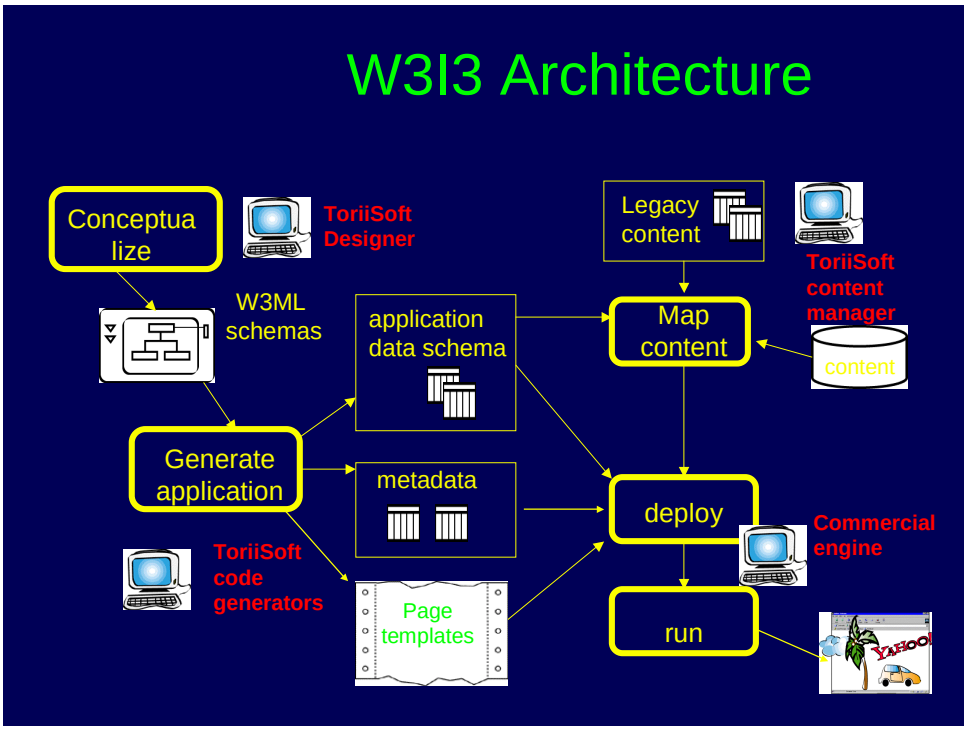
XML & Web Modelling

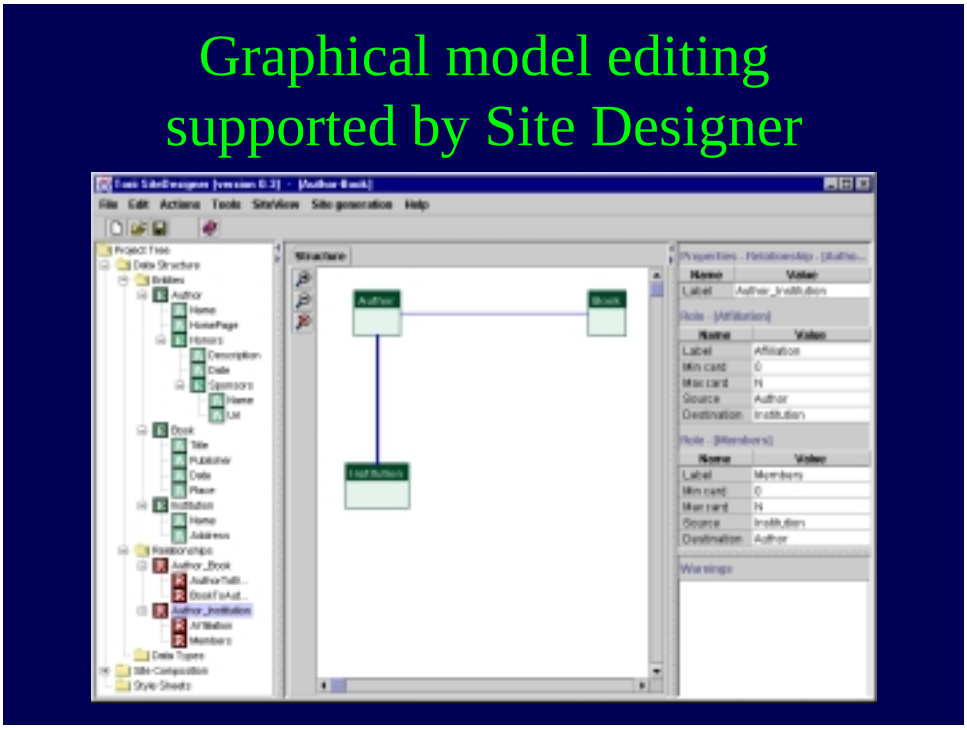
- Problem: designing data-intensive, one-to-one, multi-device Web sites with CASE
- Approach
  - **using XML for defining the syntax** of a Web Modeling Language (WebML)
  - **building XML-enabled CASE tools**
  - **using XSL for tranforming abstract XML specifications into concrete implementations (HTML+ASP, WML+Asp,HTML+JSP,….)**
  - W3I3 project: 2 years, 5 partners, 3M Ecu

W3I3 Approach

- Site = structure+composition+navigation+presentation+user



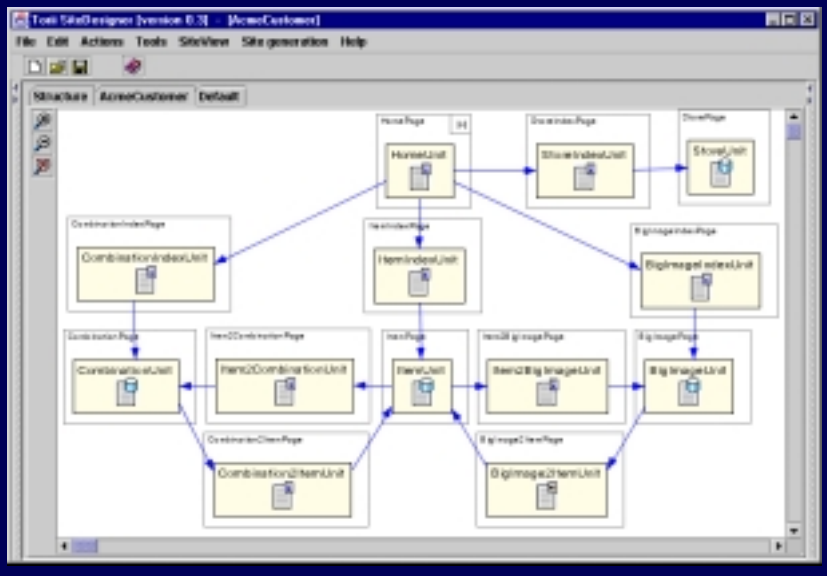

W3I3 Architecture
Conceptualize
ToriiSoft Designer
Legacy content
ToriiSoft content manager
W3ML schemas
application data schema
Map content
content
Generate application
metadata
deploy
Commercial engine
ToriiSoft code generators
Page templates
run


Graphical model editing supported by Site Designer

Example of XML Syntax

```
<ENTITY id="Author">
  <ATTRIBUTE id="Name" type="String"/>
  <ATTRIBUTE id="HomePage" type="URL"/>
  <RELATIONSHIP id="AuthorToBook" to="Book" inverse="BookToAuthor"
          minCard="1" maxCard="N"/>
  <COMPONENT id="Honors" minCard="0" maxCard="N">
   <ATTRIBUTE id="Description" type="String"/>
   <ATTRIBUTE id="DateOfHonor" type="Date"/>
   <COMPONENT id="Sponsors" minCard="1" maxCard="N">
    <ATTRIBUTE id="SponsorName" type="String"/>
    <ATTRIBUTE URL="SponsorURL" type="URL"/>
   </COMPONENT>
  </COMPONENT>
  <RELATIONSHIP id="Affiliation" to="Institution" inverse="Members"
          minCard="1" maxCard="1"/>
 </ENTITY>
```

Web site hypertext

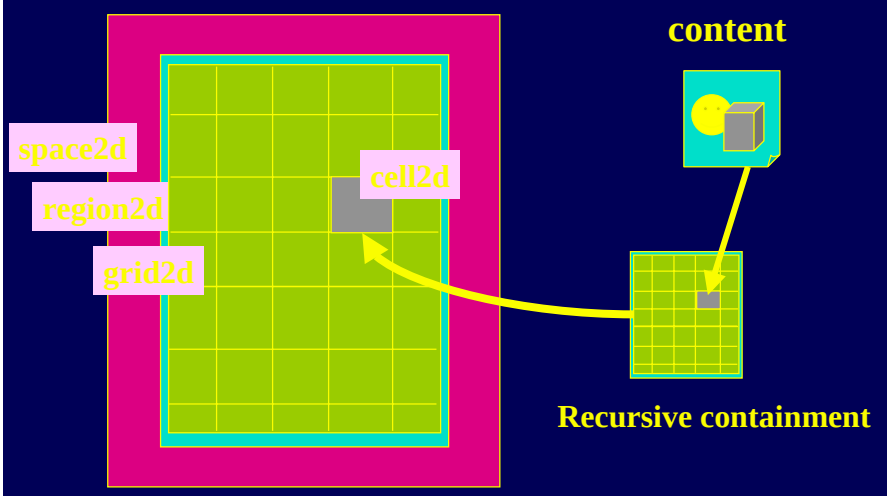


XML Syntax

```
<DATAPAGE id=”ItemData" entity=“Item">
    <INCLUDE attribute=“code"/>
    <INCLUDE attribute=“name"/>
    <INCLUDE attribute=“price"/>
    <INCLUDE attribute=“thumbnail"/>
    <LINK page=" Item2Combo" id=“l8" />
    <LINK page="DirectUnit1" id=“l12" />
</DATAPAGE>

<INDEXPAGE id="=ComboIdx" relation=“Item2Combo">
 <DESCRIPTION key=“code"/>
 <SORTATTRIBUTE name=“code" order="ascending"/>
 <LINK id=“l9" page=“ComboData" />
</INDEXPAGE>
```

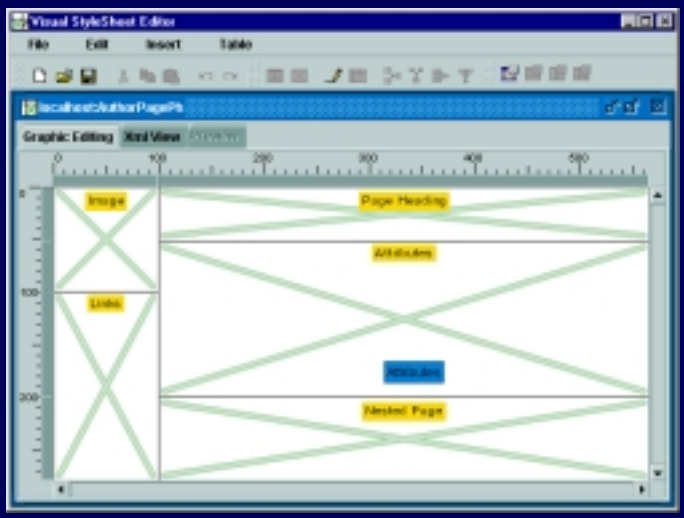
Definition of layout



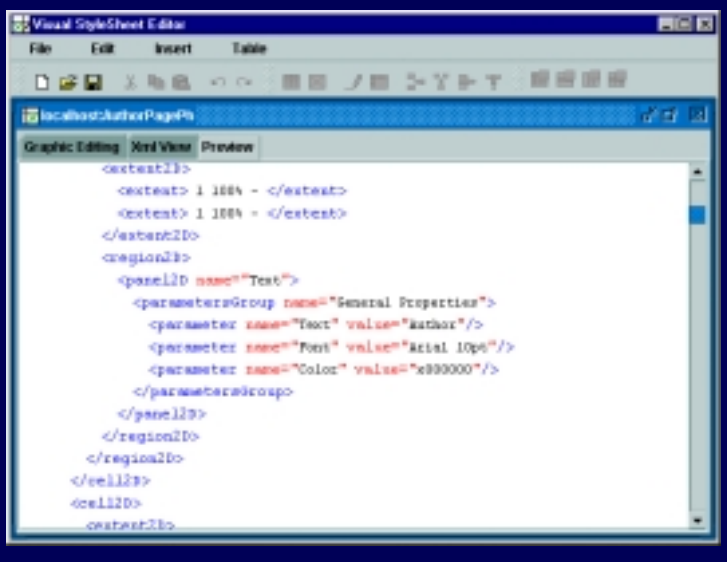
XML syntax

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE styleSheet SYSTEM "styleSheet.dtd" >
<styleSheet languages="HTML_3.2 HTML_4.0 WML" phPage="combinationPage"
    project="Acme" name="comboPage">
 <space2D>
  <region2D>
   <extent2D> <extent> 1 1024 - </extent> <extent> 1 768 - </extent> </extent2D>
   <grid2D>
    <row>    <extent> - 768 - </extent>    </row>
    <col>    <extent> - 124 - </extent>    </col>
    <col>    <extent> - 900 - </extent>    </col>
    <cell2D>
     <extent2D>  <extent> 1 1 - </extent>  <extent> 1 1 - </extent>   </extent2D>
        …. CONTENT GOES HERE……..
   </cell2D>
```

Structuring of Layout



XML Syntax



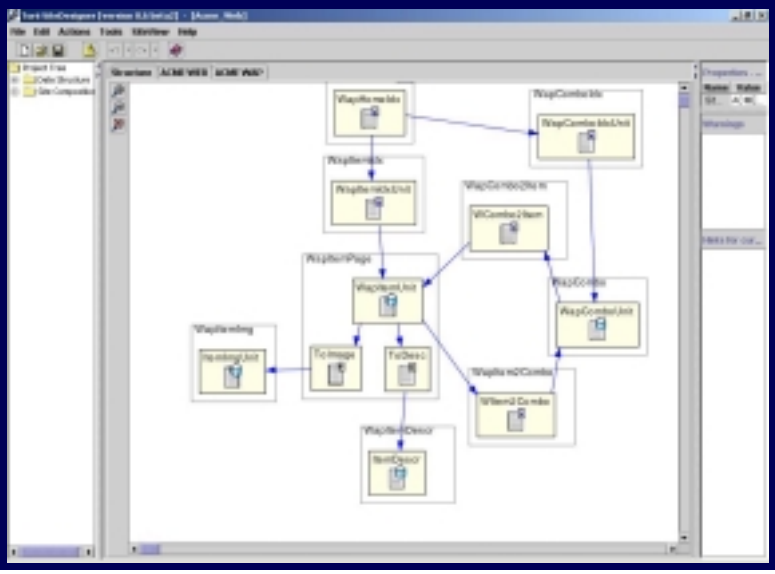
Language specific features

- Several layout or graphic features depend on the rendition language
  - e.g., useDeck for WML
- Solution
  - Panel template's attributes may be declared language-specific
  - Layout elements may be enriched with language specific attributes in a language profile

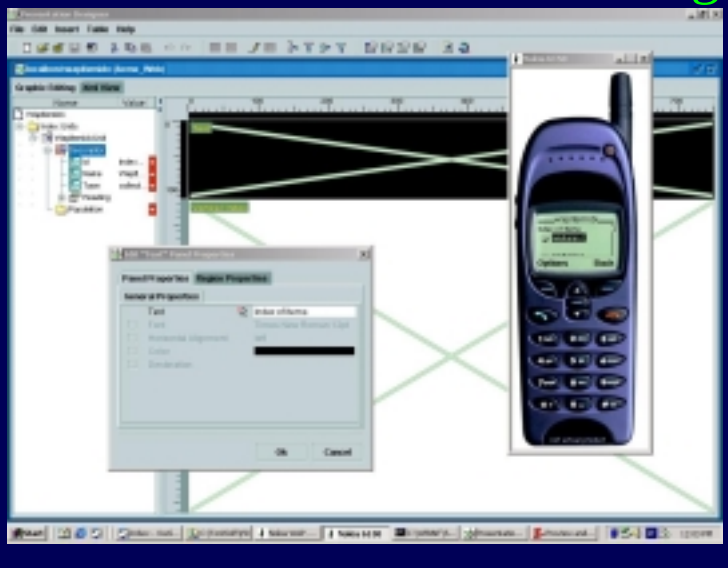
XML syntax

```
<?xml   version="1.0" encoding="UTF-8"?>
<!DOCTYPE languageProfile SYSTEM "StyleSheet.dtd">
<languageProfile name="HTML 3.2">
  <description> HyperText Markup Language 3.2& </description>
    <!-- SPACE2D ATTRIBUTES -->
    <elementAttributesDef element = "space2D">
    <attributeDef name     = "BackGround Color"  type = "Color"  presence = "implied">
      <description>  This property sets the background color  </description>
    </attributeDef>
    <attributeDef name     = "BackGround Image"  type = "Image"  presence = "implied" >
      <description>  This property sets the background image   </description>
    </attributeDef>
  </elementAttributesDef>
```

WAP site view



WAP Presentation modelling



Automatically generated WAP code (from XML, through XSL)

```
<?xml version="1.0"?>
<!DOCTYPE wml PUBLIC "-//WAPFORUM//DTD WML 1.1//EN"
"http://www.wapforum.org/DTD/wml_1.1.xml">
<!-- Source Generated by WML Deck Decoder -->
<wml>
  <card id="Page12" title="waphome">
    <do type="prev" label="Back">
      <prev/>
    </do>
    <small><img src="logoacme.gif" /></small>
    <br/>
    <small>Welcome to the ACME WAP demo site </small>
    <br/>
    <small><a href="Page13.asp"><img src="Anchor.gif" /></a></small>
    <small><a href="Page13.asp">Items </a></small>
    <br/>
    <small><a href="Page18.asp"><img src="Anchor.gif" /></a></small>
    <small><a href="Page18.asp">Combinations </a></small>
</card>
</wml>
```

For more information

www.toriisoft.com

www.webml.org

VLDB99, WWW9

Conclusions

- **Theorem: The WEB changes everything**
- **Corollary: XML is the means**

if so, the DB community has a prominent role, and should be engaged into both "fully original research" and solid transfer to the XML world of a lot of known-how