

XML

an Introduction

G. Falquet
L. Nerima

ISI
© G. Falquet, L. Nerima, CUI - Université de Genève 1

Different perspectives on XML

- Document engineering
- Web
- Information interchange
- Database
- Domaine specific standards and applications

ISI
© G. Falquet, L. Nerima, CUI - Université de Genève 2

XML - Document Markup View

Markup language
Markup tags in documents
At different levels

formatting tags

... please read this ***bold***BEFORE*plain* installing your ...

semantic tags

//issue// {Should we use a standard language ? }
//position//{Not in this situation} //argument// {There is no ...}

ISI
© G. Falquet, L. Nerima, CUI - Université de Genève 3

Example - HTML

Direct formatting tags

please read the following section

Indirect formatting (through style sheets)

<H1>Consistency checking</H1>

Content description

<ADDRESS> A. Zweistein, CERN </ADDRESS>

ISI
© G. Falquet, L. Nerima, CUI - Université de Genève 4

XML Markup

XML markup can be used at all levels
(in different documents)

```
...
<frame width="3.2cm" height="6.5cm">
  New highlights ...
</frame>
...
```

```
<technical-report>
  <author>Me</author>
  <keyword>rockets</keyword>
  <keyword>taxes</keyword>
  ...
</technical-report>
```

5

XML Genericity

XML is not a language
It is a meta-language / language generator

You must define your own language

your own tags

your interpretation of the tags

how these tags must be processed / transformed

© G. Falquet, L. Nerima, CUI - Université de Genève

6

The Ancestor: SGML

The idea is not new

SGML : Standard Generalized Markup Language

Charles Goldfarb (IBM) --> GML
normalized by ISO - 1986

Principle: tags to describe the logical structure of a documents

Aim: electronic document publication and exchange

Separate content and presentation (independence)

© G. Falquet, L. Nerima, CUI - Université de Genève

7

General structure of a document

An XML document is made of a prologue and **elements**

Elements contain

sequences of characters

other elements

nothing

An element can have attributes

attribute name = value

There is a "root" element

© G. Falquet, L. Nerima, CUI - Université de Genève

8

An XML Document

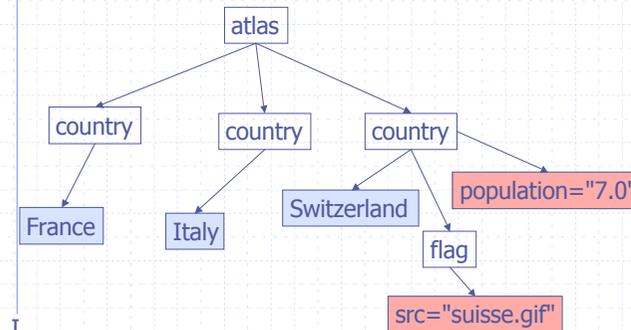
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<atlas>
  <country> France </country>
  <country> Italy </country>
  <country population="7.0"> Switzerland
    <flag src="suisse.gif"/>
  </country>
</atlas>
```

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

9

Hierarchical view



ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

10

Formal structure of an XML Document

XMLDocument =

Prolog
Element

Prolog = <?xml version=V encoding=E ?>

V = "1.0"

E = "UTF-8" or "ISO-8859-1" or ...

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

11

Elements

element = <Name Attributes> content </Name>

content = (repeated)

element | CharData | Reference | CDSection |
ProcessingInstruction | Comment

Attributes = (repeated)

Name = value

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

12

Elements - examples

```
<city>Paris</city>
<item>
  <category>food</category>
  <desc>A special cake made of ... </desc>
  <price>SFr. 56.30</price>
</item>

<paragraph>This system <note>as of March 2001</note>
is comprised of 22 modules</paragraph>
```

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

13

Elements - examples with attributes

```
<item category="food" price="SFr. 56.30">
  <desc>A special cake made of ... </desc>
  <evaluation>no toxic element found</evaluation>
</item>
```

A data modelling problem: attribute or element ?
(similar to "class or attribute" in UML)

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

14

Elements - examples with references

```
<paragraph>
  This processor was developed by &MIT; and &CMU; to
  ...
  <!-- Entity references are defined elsewhere -->
</paragraph>
<important>
  The less-than symbol must be replaced by &lt; , while
  other (Unicode) characters like &#2341 or &#x3CC6
  may be written in decimal or hexadecimal code.
</important>
```

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

15

Elements with special characters

```
<a-complex-description>
  <![CDATA[ Here you can write < and > or [ and quotes
  like this " etc. etc.
  the end is marked by ]]>
</a-complex-description>
```

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

16

Declarations

An XML document may contain declarations

```
Prolog =  <?xml version=V encoding=E ?>
          <!DOCTYPE Name [
            declarations
            ... ]
          >
```

declaration

to declare entities

to declare a document type (grammar)

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

17

Entity declarations

```
<?xml version=V encoding=E ?>
<!DOCTYPE report [
  <!ENTITY CUI "Centre universitaire d'informatique">
  <!ENTITY UN "United Nations">
] >
<report> ... &UN; ... &CUI; ...
```

Predefined entities:

< , > , & , ' , " , "

Also, external text or binary entities

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

18

Document syntax declaration

```
<!DOCTYPE report [
  <!ELEMENT name decl>
  <!ATTLIST name decl>
  ...
] >
<report> .... </report>
```

Defines the content of each element (language)

Documents must be valid (belong to the specified language)

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

19

External document type definition

```
<!DOCTYPE report
  SYSTEM "http://a.b.ch/dtds/report.dtd">
<report>
  ....
</report>
```

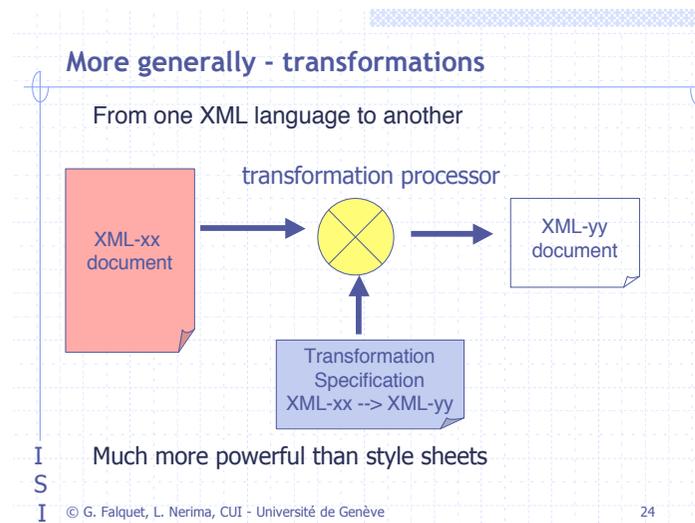
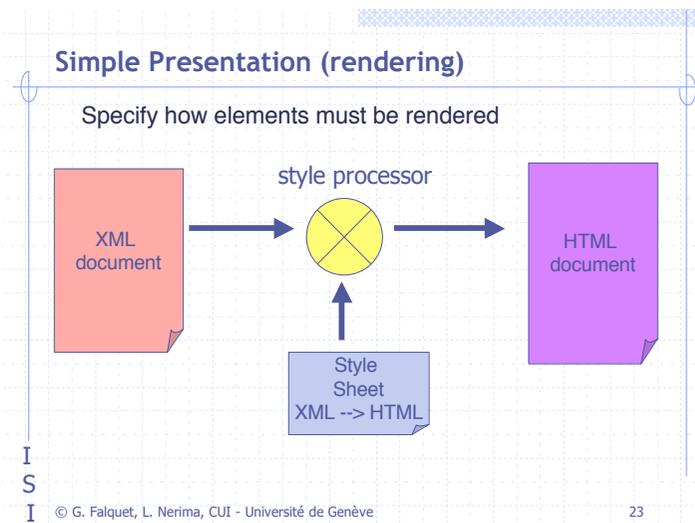
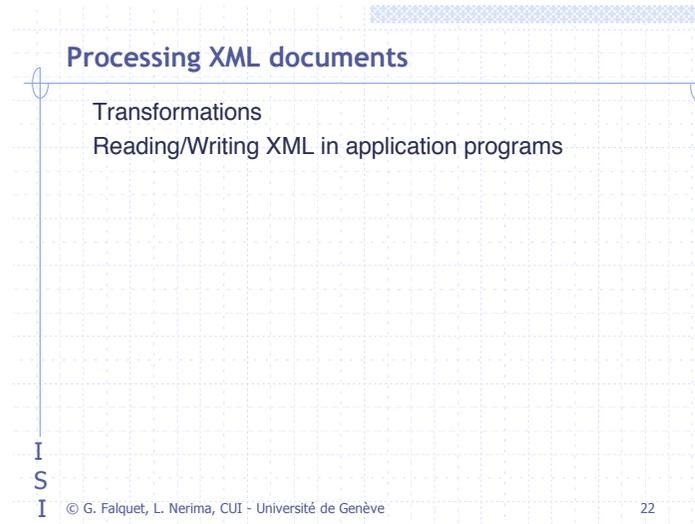
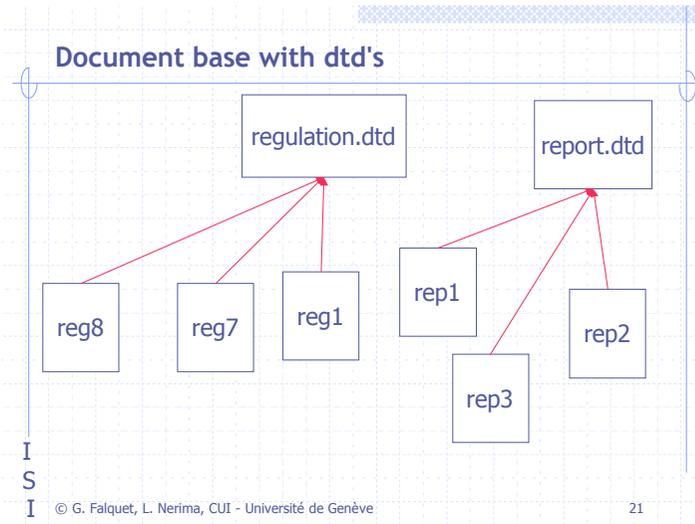
Unique (centralized) definition

Easier to maintain and evolve

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

20



Example

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<?xml-stylesheet type="text/xsl" href="pays.xsl" title="1"?>
<atlas xmlns:ISI="http://cuiwww.unige.ch/isi">
  <pays >
    <nom> France </nom>
    <capitale> Paris </capitale>
    <langue>FR</langue>
  </pays>
  <pays >
    <nom> Austria </nom>
    <capitale> Vienna </capitale>
    <langue>DE</langue>
  </pays>
  <pays >
    <nom> Switzerland </nom>
    <capitale> Bern </capitale>
  </pays>
</atlas>
```

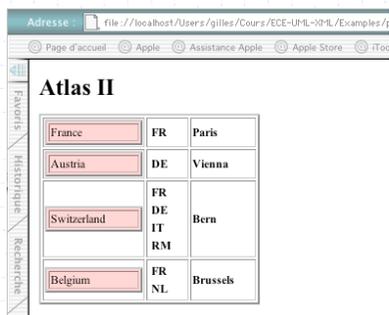
I
S
I

Transformation 1



I
S
I

Transformation 2

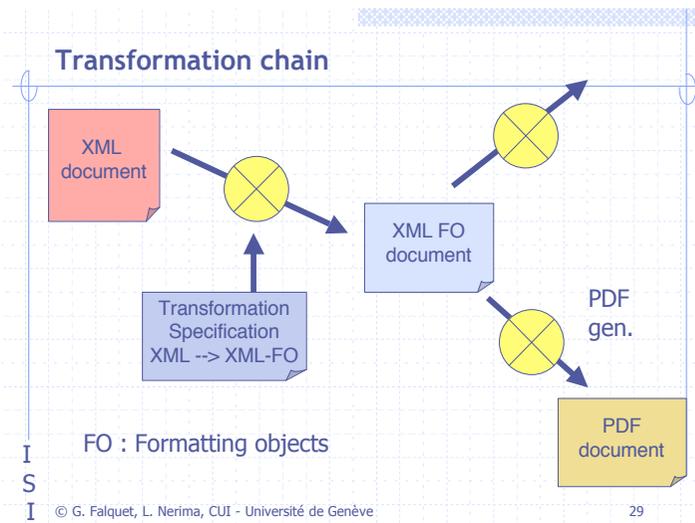


I
S
I

Transformation 3



I
S
I



Reading/Writing XML in applications

XML input must be parsed

Two types of processing

- Sequential (event driven), elements parsed one at a time
- Tree
 - the parsing process produces a tree (in memory)
 - the application works on this tree (tree manipulation)

Standard Application Programming Interface

Simple API for XML (SAX 1.0)

© G. Falquet, L. Nerima, CUI - Université de Genève 30

XML Applications (Languages)

Many XML language are being defined
Some have been standardized

- MathML
- SVG
- X3D
- XMI

© G. Falquet, L. Nerima, CUI - Université de Genève 31

Existing languages - MathML

MathML: mathematical markup

decompose equations in basic components (numbers, constants, operations, ...)

```

<apply>
  <power/>
  <apply>
    <plus/>
    <ci>a</ci>
    <ci>b</ci>
  </apply>
  <cn>2</cn>
</apply>
    
```

represents $(a+b)^2$

© G. Falquet, L. Nerima, CUI - Université de Genève 32

Language : SMIL

Synchronized Multimedia Integration Language

Describe how to play MM contents
sequentially, in parallel, with alternatives
duration, position on screen, ...

```
<seq>
  <item>
    
  </time>
</par>
<item>
  ...
</seq>
```

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

33

Language - SVG : Scalable Vector Graphics

Vector Graphics ≠ Bitmap images (GIF, PNG, JPEG, ...)

Can be viewed at any scale

```
<?xml version="1.0" standalone="no"?>
<!DOCTYPE svg PUBLIC "-//W3C//DTD SVG 20010904//EN"
"http://www.w3.org/TR/2001/REC-SVG-20010904/DTD/svg10.dtd">
...
<ellipse cx="-439" cy="219" rx="27" ry="18"
  style="fill:#73bf73; stroke:rgb(66,105,66); fill-opacity:100%; "
  id="ellipse113"
  transform="matrix(1 0 0 1 992.753 141.329)">
</ellipse>
<path
  d="M-400 152C-406 155 -417 162 -425 171 -427 173 -429 183 -434 202"
```

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

34

Language X3D

3-dimensional scenes (virtual reality scenes)

Can be transformed into VRML
VRML viewers for Web Browsers

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

35

Language XML

To represent UML models in XML

UML concept ---> XML tag

To exchange UML models

A standardized way to store UML models
independent of the UML tool makers

ISI

© G. Falquet, L. Nerima, CUI - Université de Genève

36

Your own language ...

There are many XML languages in

- banking
- e-commerce
- science
- libraries
- etc.

But you probably have specific needs ...

I
S
I

© G. Falquet, L. Nerima, CUI - Université de Genève

37

XML and the Web

An important component for the semantic Web

- add semantics to web pages
- HTML is presentation-oriented

- semantic search
- "find Paris as a street name, not the city, not a person name, ..."

- use web sites as databases
- publish and exchange data on the web

I
S
I

© G. Falquet, L. Nerima, CUI - Université de Genève

38

XML and Databases

A (not so) new database model

Semi-structured data

- less rigid than relational schemas
- model less structured objects

Integrate document bases and databases

XQL : XML query language

Native XML database systems

- but Oracle already makes plans (and tools) to kill them

I
S
I

© G. Falquet, L. Nerima, CUI - Université de Genève

39