Enabling Grids for E-science in Europe

www.eu-egee.org

3 – 4 June 2004

Introduction to XML

EGEE is a project funded by the European Union under contract IST-2003-508833

Objectives

Enabling Grids for E-science in Europe

- To understand basic XML syntax
- To explore the concept of namespaces
- To understand the role of Schema

What is XML

Enabling Grids for E-science in Europe

- XML stands for extensible markup language
- It is a hierarchical data description language
- It is a sub set of SGML a general document markup language designed for the American millitary.
- It is defined by w3c.

How does XML differ from HTML?

Enabling Grids for E-science in Europe

- HTML is a presentation markup language – provides no information about content.
- There is only one standard definition of all of the tags used in HTML.
- XML can define both presentation style and give information about content.
- XML relies on custom documents defining the meaning of tags.

What is a Schema?

Enabling Grids for E-science in Europe

- A schema is the definition of the meaning of each of the tags within a XML document.

- *Analogy: A HTML style sheet can be seen as a limited schema which only specifies the presentational style of HTML which refers to it.*

- *Example: in HTML the tag pre-defined. In XML you would need to define this in the context of your document.*

Pre-existing schema

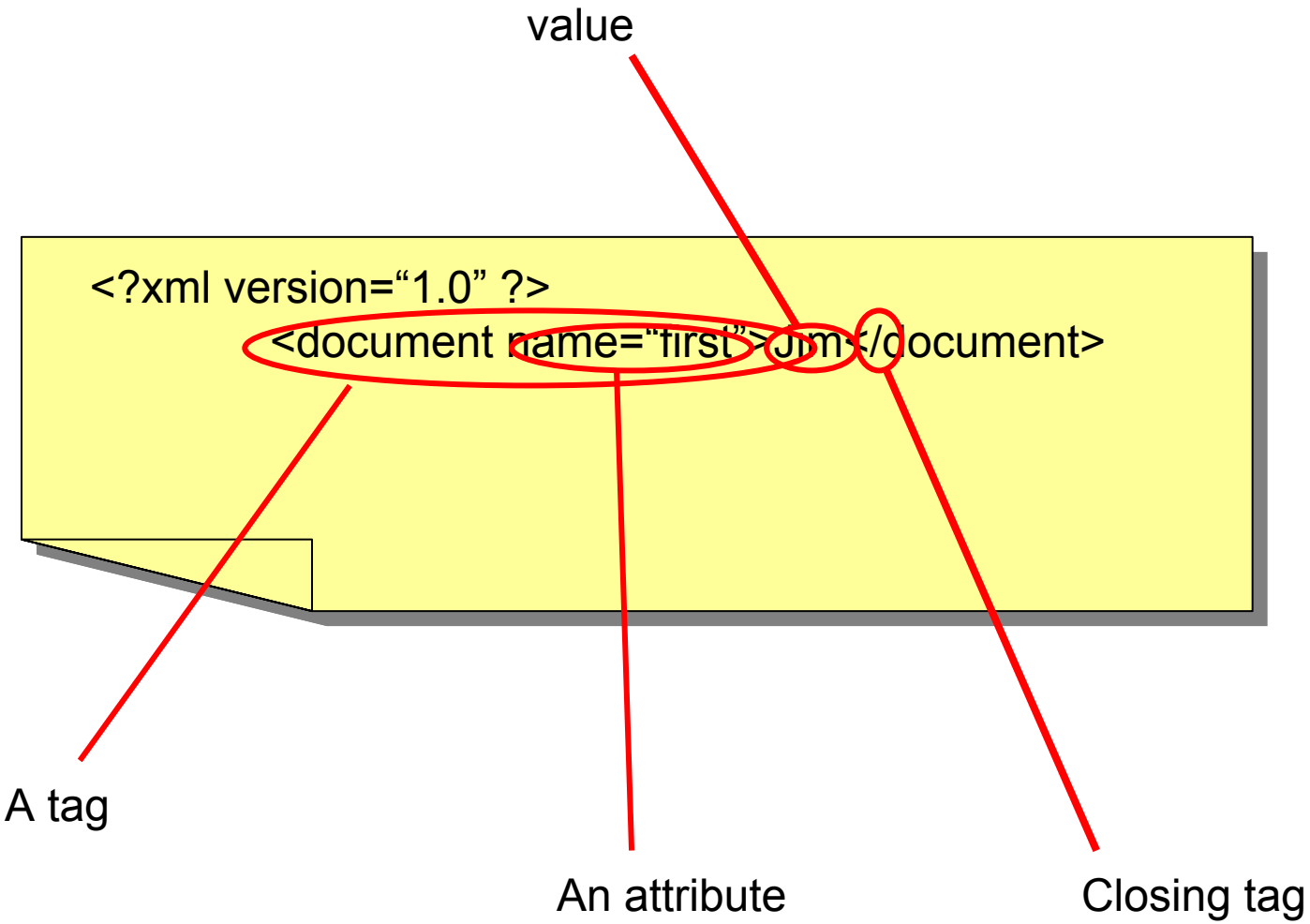
Enabling Grids for E-science in Europe

- A schema can ‘inherit’ from another and extend it. (*analogous to extending a class in JAVA*)

- For example the basic tags which allow you to write schema are defined in :

  http://www.w3.org/2001/XMLSchema

A minimal XML document



Valid and well formed

Enabling Grids for E-science in Europe

- A correct XML document must be both valid and well formed.
- Well formed means that the syntax must be correct and all tags must close correctly (eg <…> </…>).
- Valid means that the document must conform to some XML definition (a DTD or Schema).

(Otherwise there can be no definition of what the tags mean)

Namespaces in XML

Enabling Grids for E-science in Europe

- Schema require namespaces.
- A namespace is the domain of possible names for an entity within a document.
- Normally a single namespace is defined for a document. In this case fully qualified names are not required.

Common namespace prefixes

Enabling Grids for E-science in Europe

xsi	http://www.w3c.org/2000/10/XMLSchema-instance *namespace governing XMLSchema instances*
xsd	http://www.w3c.org/2000/10/XMLSchema *namespace of schema governing XMLSchema (.xsd) files*
tns	by convention this refers to “this” document *refers to the current XML document*wsdl http://schemas.xmlsoap.org/wsdl/ *WSDL namespace*
soap	http://schema.xmlsoap.org/wsdl/soap/ *WSDL SOAP binding namespace*

Using namespaces in XML

Enabling Grids for E-science in Europe

- To fully qualify a namespace in XML write the namespace:tag name. eg.

  <my_namespace:tag> </my_namespace:tag>

- In a globally declared single namespace the qualifier may be omitted.

- More than one namespace:

  <my_namespace:tag> </my_namespace:tag>

  <your_namespace:tag> </your_namespace:tag>

  can co-exist if correctly qualified.

Namespaces in programming languages

Enabling Grids for E-science in Europe

- In C/C++ defined by #includes and classes (eg. myclass::variable).
- In PERL defined by package namespace, $local and $my (eg. myPackage::variable).
- In JAVA defined by includes and package namespace (eg. java.lang.Object)
- **Defines the scope of variables**

Why namespaces in XML?

Enabling Grids for E-science in Europe

- A namespace is used to ensure that a tag (variable) has a unique name and can be referred to unambiguously.
- Namespaces protect variables from being inappropriately accessed – encapsulation.
- This makes sure that when you access a variable correctly it has the expected value.

Schema

```
<?xml version="1.0"?>
<xs:schema xmlns:xs=http://www.w3.org/2001/XMLSchema
xmlns="document" >
<xs:element name = "DOCUMENT">
                <xs:element name="CUSTOMER"> </xs:element>
</xs:element>
</xs:schema>
```

Simple schema saved as order.xsd

```
<?xml version="1.0"?>
<DOCUMENT xmlns="document"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
Xsi:schemaLocation="order.xsd">
<DOCUMENT>
                <CUSTOMER>sam smith</CUSTOMER>
                <CUSTOMER>sam smith</CUSTOMER>
</DOCUMENT>
```

XML document derived from schema.

Document Type Definition (DTD)

```
<?xml version="1.0">
<!DOCTYPE DOCUMENT [
<!ELEMENT DOCUMENT (CUSTOMER)>
<!ELEMENT CUSTOMER (#PCDATA)>
]>
```

Simple DTD saved as order.dtd

```
<?xml version="1.0"?>
<!DOCTYPE DOCUMENT SYSTEM "order.dtd">
<DOCUMENT>
        <CUSTOMER>sam smith</CUSTOMER>
        <CUSTOMER>sam smith</CUSTOMER>
</DOCUMENT>
```

XML document derived from DTD.

URI vs URL

Enabling Grids for E-science in Europe

- This is similar to the distinction between an class and an instance in Object Oriented Programming.
- A URI is a universal resource identifier which could have many forms (ie could be an ISBN number if these were in a URN scheme)
- A URL is a http instance of a URI
- URN (universal resource name) is the declared name of a resource
- (URC {citation} would point to metadata

Areas of XML Application

Enabling Grids for E-science in Europe

- Document Definition
- Data Exchange
- Metadata (Data about Data)
- Remote Procedure Calls

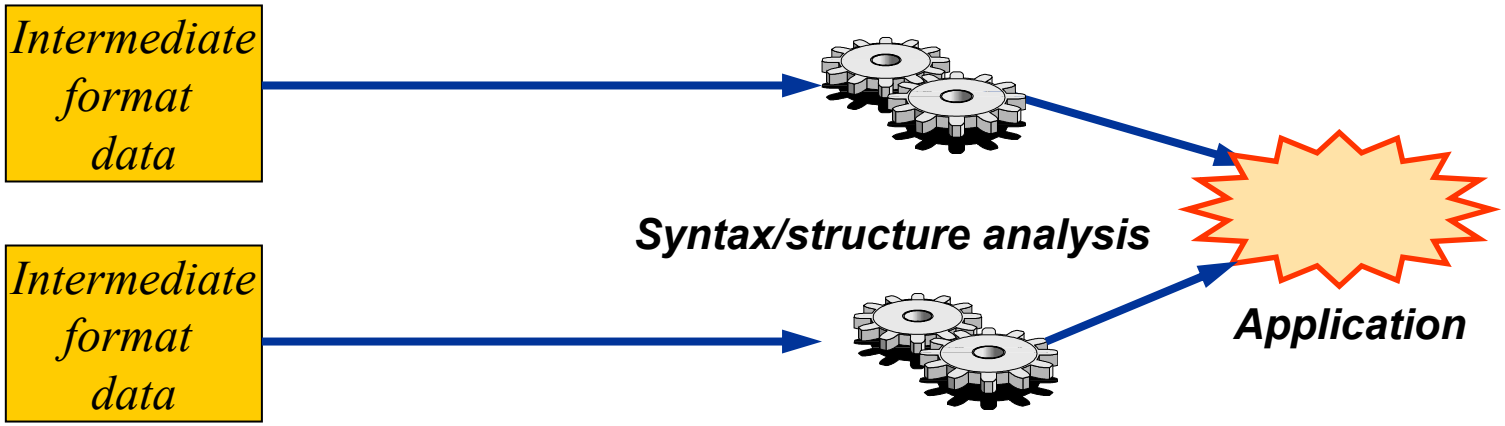
Document Definition

Enabling Grids for E-science in Europe

- XML used in particular applications – SGML users
  - Specialised XML Editors
- Word2000 uses XML/HTML hybrid, all OS X applications use XML configuration files.
- Microsoft .NET initiative
  - Documents encoded in XML
  - Information providers expose data in XML
  - More widespread tools (MS Word?)

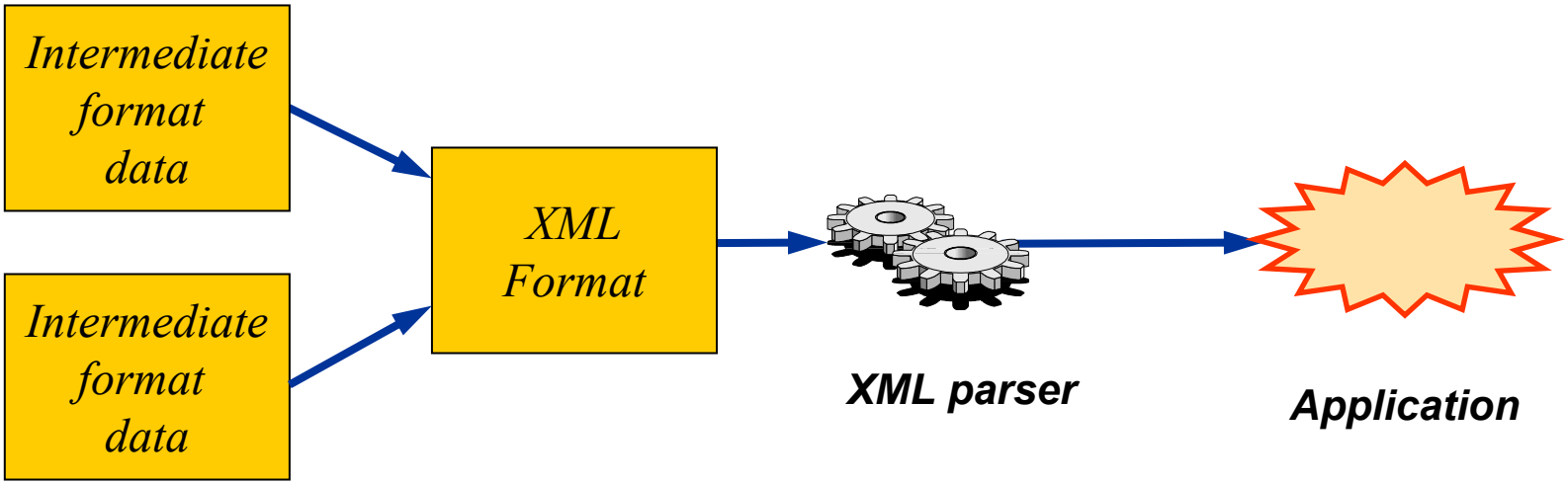
Using XML for Data Exchange - Current

- Many applications express their data in an intermediate format, to aid interoperability with other applications
  - Other applications parse these documents to reconstitute the data



Using XML for Data Exchange - Future

- XML can help, because its (standard) notation can be analysed by off-the-shelf XML parsers



Using XML as Metadata

Enabling Grids for E-science in Europe

- XML metadata provides information about the structure and meaning of any data
- XML metadata can be used to perform more intelligent web searches for goods or information
- Cross-site searches are difficult (depends on metadata info in pages)
- XML metadata is more self-describing and meaningful, for example ...
  - Search for all plays written by William Shakespeare
  - Rather than every web site that mentions him!

Using XML for Remote Procedure Calls

Enabling Grids for E-science in Europe

- XML used to exchange data between Software Components
- Simple Object Access Protocol – SOAP
  - A lightweight protocol for exchange of information in a decentralised, distributed environment
  - Web-Sites expose interfaces for interrogation
- Universal Description, Discovery and Integration – UDDI
  - Integrating business services
  - 'Yellow/White Pages'

Support for XML

Enabling Grids for E-science in Europe

- Driven by World Wide Consortium (W3C)
- Industry bodies (OASIS, BizTalk)
- Microsoft, Sun, Oracle, IBM, Novell…
- Dell – large implementation of XML
- Inland Revenue - eGIF

Industry perspectives

Enabling Grids for E-science in Europe

"I believe both Microsoft and the industry should really bet their future around XML, the standards around XML are key to where we need to go."

Bill Gates, Microsoft

"XML has the potential to address some of the traditional failings of message standards. Its impact could be considerable."

Bank of England

Use of XML in biological databases

Enabling Grids for E-science in Europe

- EBI Molecular Structure Database (MSD) is an extraction from PDB (Protein Data Bank) which is encoded in XML.
- Uses DTDs
- Initiatives at EBI, NCBI and else where to use XML to make heterogeneous databases interoperable

Summary

Enabling Grids for E-science in Europe

- XML is a language that provides
  - A mark-up specification for creating self descriptive data
  - A platform and application independent data format
  - A way to perform validation on the structure of data
  - A syntax that can be understood by computers and humans
  - The way to advance web applications used for electronic commerce.