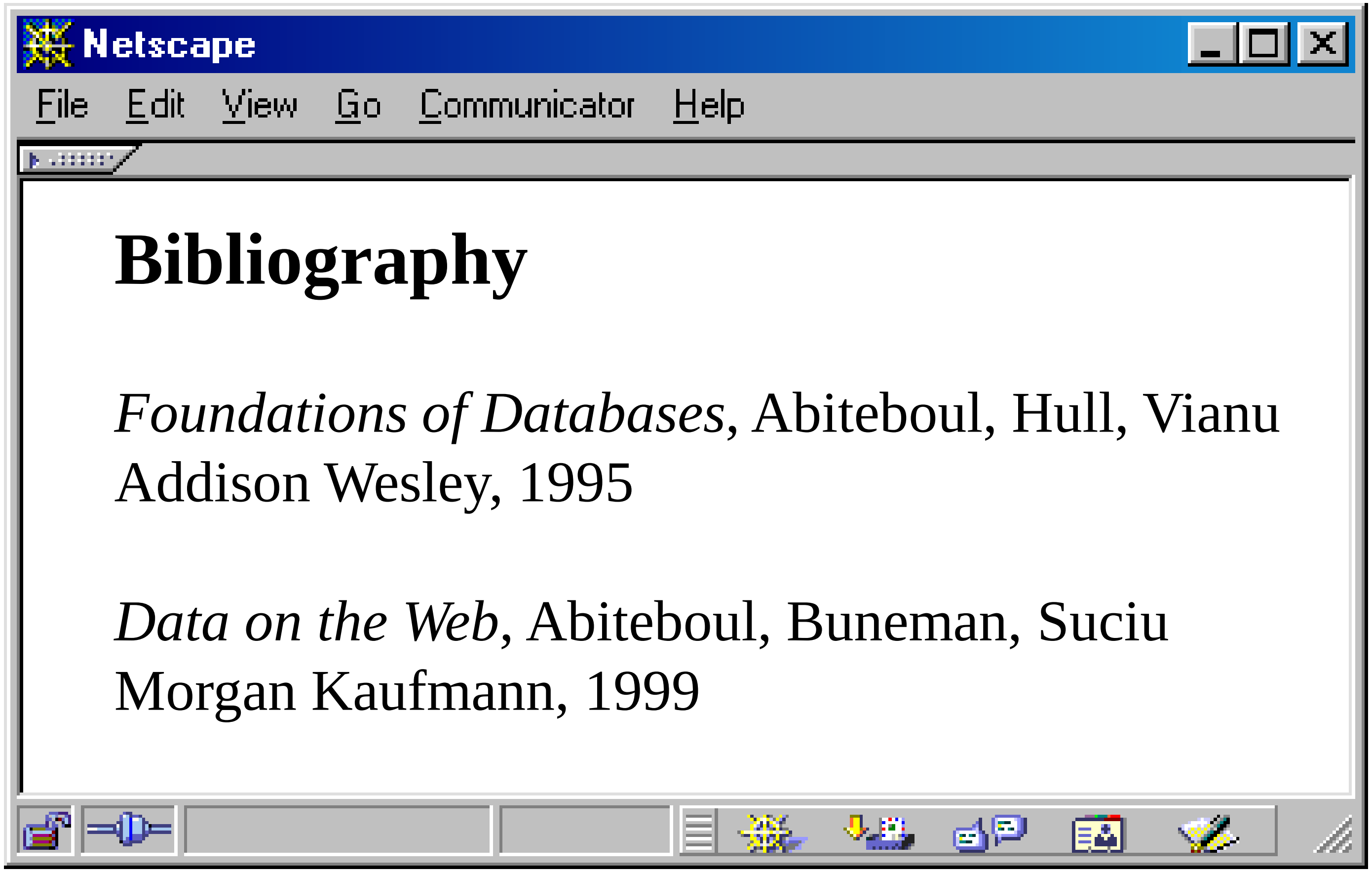
Introduction to XML

- XML stands for EXtensible Markup Language
- XML is a markup language for documents containing semistructured information.
- a W3C (The World Wide Web Consortium) standard to complement HTML.
- In HTML, both the tag semantics and the tag set are fixed.
- XML tags are not predefined. Users must define their own tags.
- Origins: structured text SGML
- SGML is the Standard Generalized Markup Language defined by ISO (International Organization for Standardization), but it is not well suited to serving documents over the web.

- motivation:
  - HTML describes presentation format
  - XML describes content
  - XML is not a replacement for HTML
- XML $\subset$ SGML
- XML documents use a self-describing and simple syntax.
- XML can also use a Document Type Definition (DTD) or an XML Schema to describe the structure of data (optional).
- http://www.w3.org

From HTML to XML



Bibliography

Foundations of Databases, Abiteboul, Hull, Vianu
Addison Wesley, 1995

Data on the Web, Abiteboul, Buneman, Suciu
Morgan Kaufmann, 1999

HTML describes the presentation

HTML

```
<h1> Bibliography </h1>
<p> <i> Foundations of Databases </i>
        Abiteboul, Hull, Vianu
        <br> Addison Wesley, 1995
<p> <i> Data on the Web </i>
        Abiteboul, Buneman, Suciu
        <br> Morgan Kaufmann, 1999
```

XML

```
<bibliography>
    <book>  <title> Foundations… </title>
            <author> Abiteboul </author>
            <author> Hull </author>
            <author> Vianu </author>
            <publisher> Addison Wesley </publisher>
            <year>  1995 </year>
    </book>

    …
</bibliography>
```

Note: XML describes the content.

Tag names provide some meanings. They are defined by people writing the XML document.

XML Terminology

- tags: book, title, author, ...
- start tag: **<book>**
- end tag: **</book>**
- elements: **<book>** ... **</book>**,
  <author> ... **</author>**
- elements can be nested
- empty element: **<red> </red>** abbrev. **<red/>**
- an XML document: a single *root* element

Well formed XML document: if tags are properly nested and attributes of an element are unique.

More XML: Attributes

```
<book price = “55” currency = “USD”>
  <title> Foundations of Databases </title>
  <author> Abiteboul </author>
  …
  <year> 1995 </year>
</book>
```

- Attributes (e.g. price and currency) are alternative ways to represent data.

- Design issue: When should data be designed as an element **or** an attribute? This is a big problem of XML.

More XML: Attributes *vs.* Elements

Q: When should data be designed as an element or an attribute of an element?

E.g. We could represent the information about Alan as

```
<person> <name> Alan </name>
         <age> 42 </age>
         <email> agb@abc.com </email>
</person>
```

or

```
<person name=“Alan”  age=“42”  email=“agb@abc.com”/>
```

or

```
<person age=“42”>
       <name> Alan </name>
        <email> agb@abc.com </email>
</person>
```

Q: Which way is the best?

More XML: oids and References

```
<person pid=“o555”> <name> Jane </name> </person>

<person pid=“o456”> <name> Mary </name>
   <children cpids=“o123 o555”/>
</person>

<person pid=“o123” mother=“o456”> <name> John </name>
</person>
```

- oids and references in XML are just syntax
- Problem: redundancy
- All attributes are single valued attributes except IDREFS type of attributes (e.g. the attribute children - cpids). See DTD lecture notes.

More XML: Order

- Order for elements in XML documents are important. However, attributes are not ordered in XML.

E.g. The following 2 XML documents are different, not equivalent (because of order of elements):

(1) <person><firstname> John </firstname>
<lastname> Smith </lastname> </person>

(2) <person><lastname> Smith </lastname>
<firstname> John </firstname> </person>

However, the following 2 XML documents are equivalent.

(1) <person firstname=“John” lastname=“Smith”/>

(2) <person lastname=“Smith” firstname=“John”/>

Note: XML Schema now allows child elements to be declared as unordered.
(see lecture notes on XML Schema)

More XML: Mixing Elements and Text

- XML allows us to mix PCDATA (Parsed Character Data, i.e.text) and subelements (child elements) within an element.

```
<person>
  This is my best friend
  <name> Alan </name>
  <age> 42 </age>
  I am not too sure of the following email
  <email> agb@abc.com </email>
</person>
```

Note: Bad format and design!