An Introduction to XML

Emmanuel Stefanakis

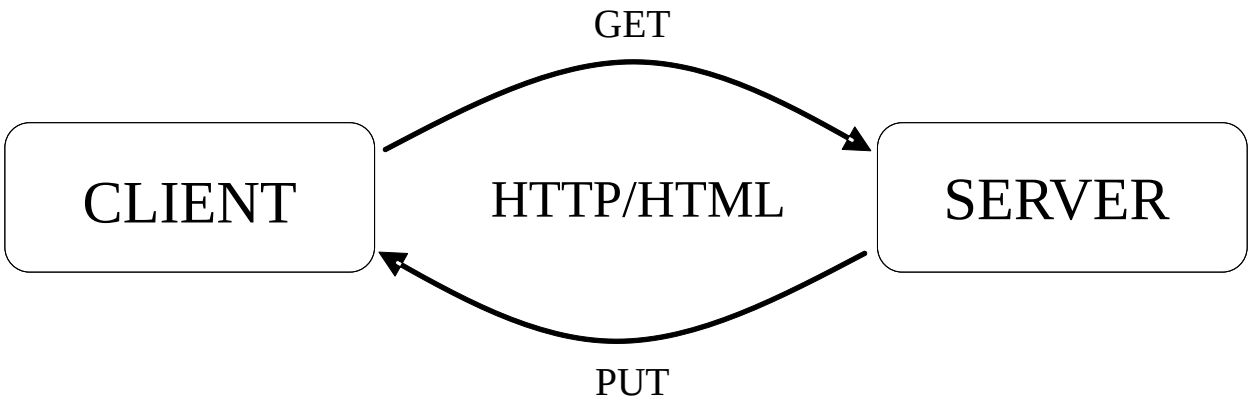
estef@unb.ca

Introduction

- eXtensible Markup Language (XML) …
  - is a widely accepted format for describing and exchanging data.
- The slides provide …
  - a brief overview of the XML technology, and its relationship to database technology
- Outline …
  - XML and Data
  - XML Infrastructure
  - XML Family of Technologies
  - XML and Databases
  - Repositories for XML Documents

Introduction

- Air Canada → Online reservation

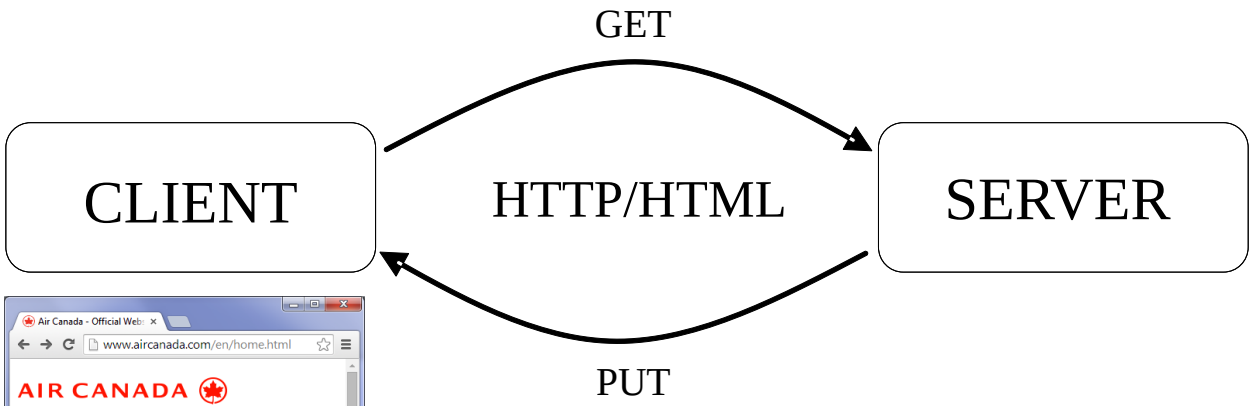


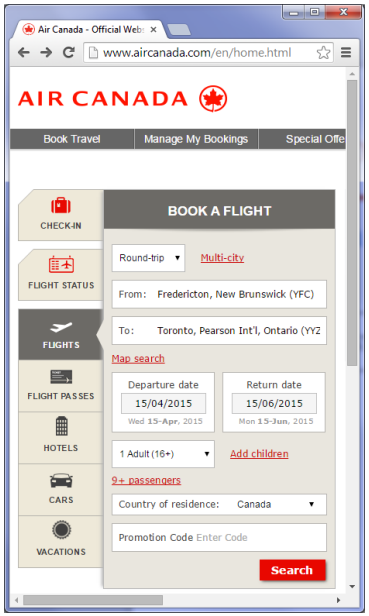
you...

AIR CANADA

Introduction

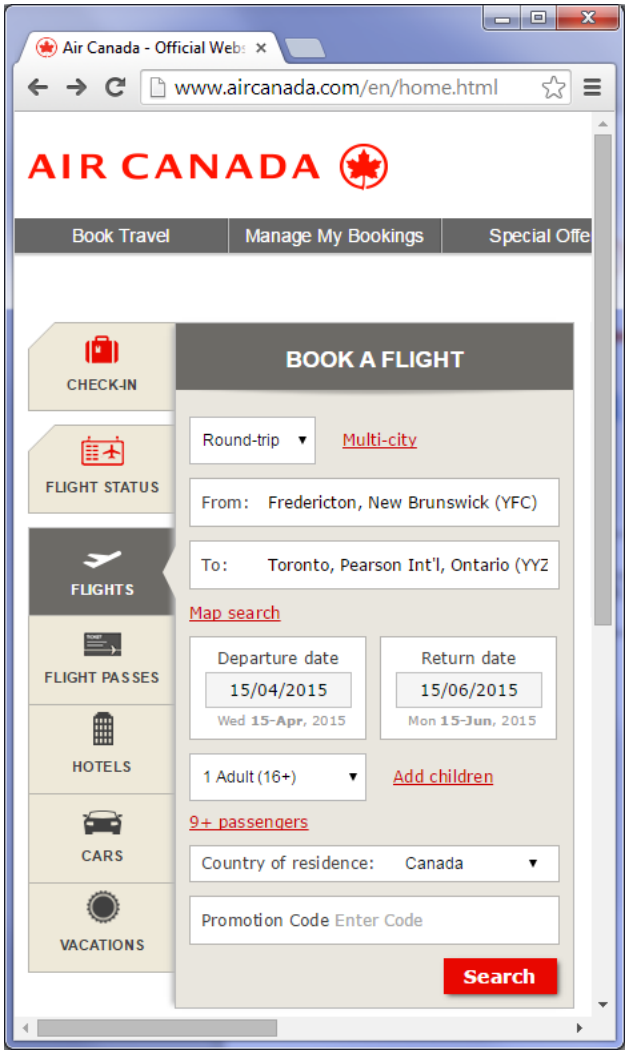
- Air Canada → Online reservation





AIR CANADA

Introduction



HTML page…

Type: Round trip

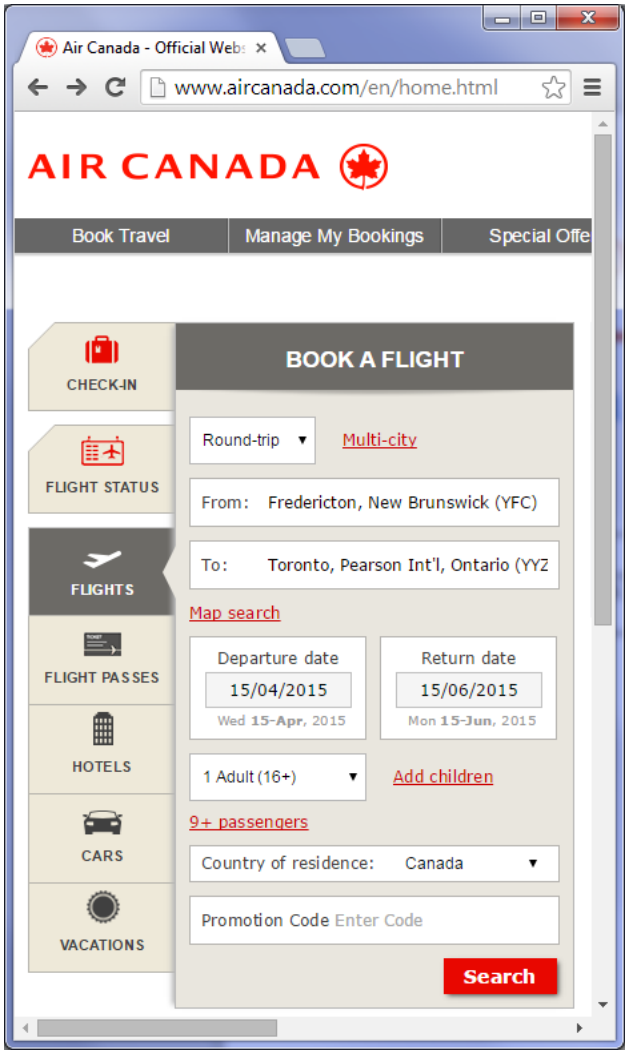
From: Fredericton

To: Toronto

Departure: April 15, 2015

Return: June 15, 2015

Introduction



HTML page…

Data

Type: Round trip

From: Fredericton

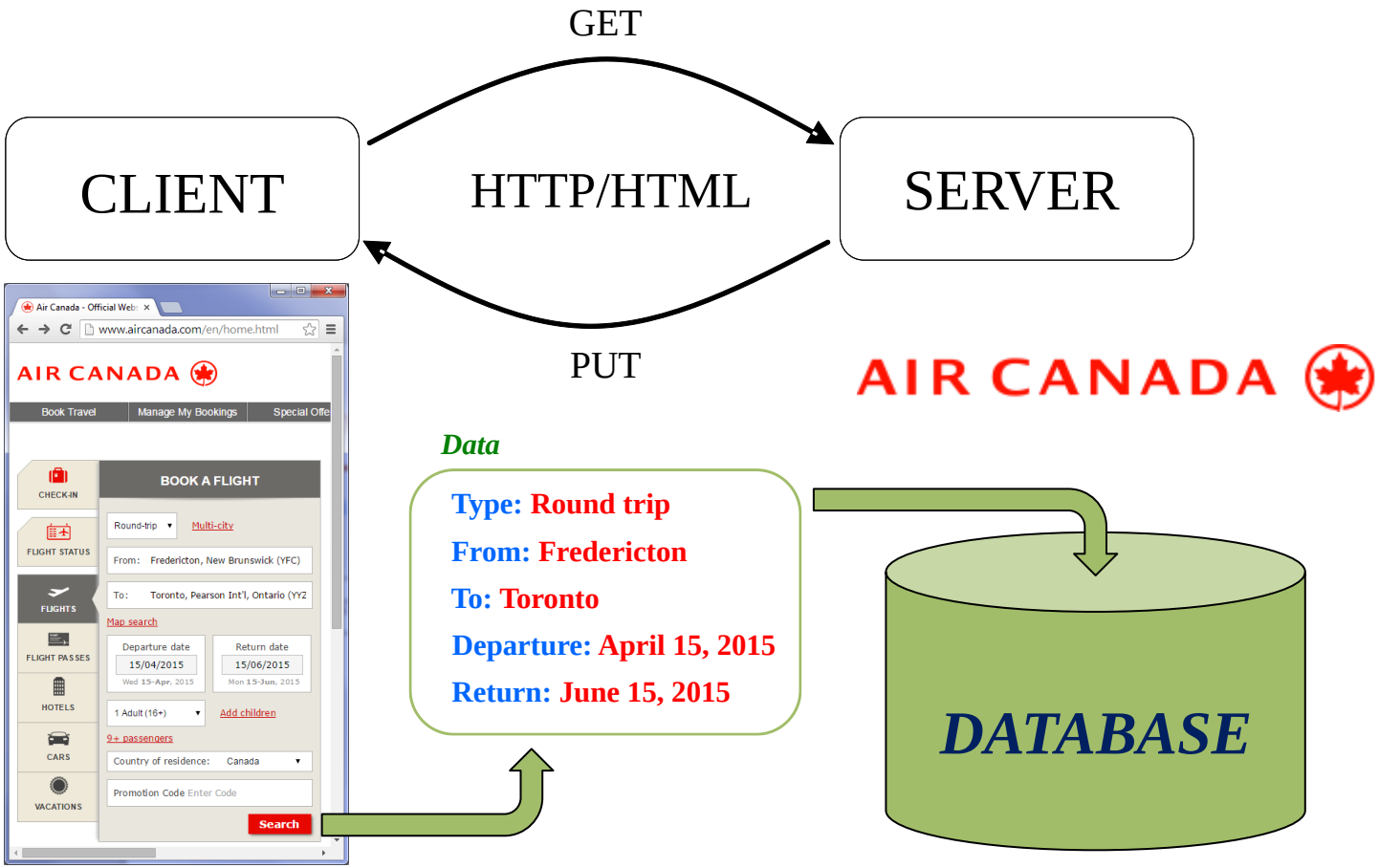
To: Toronto

Departure: April 15, 2015

Return: June 15, 2015

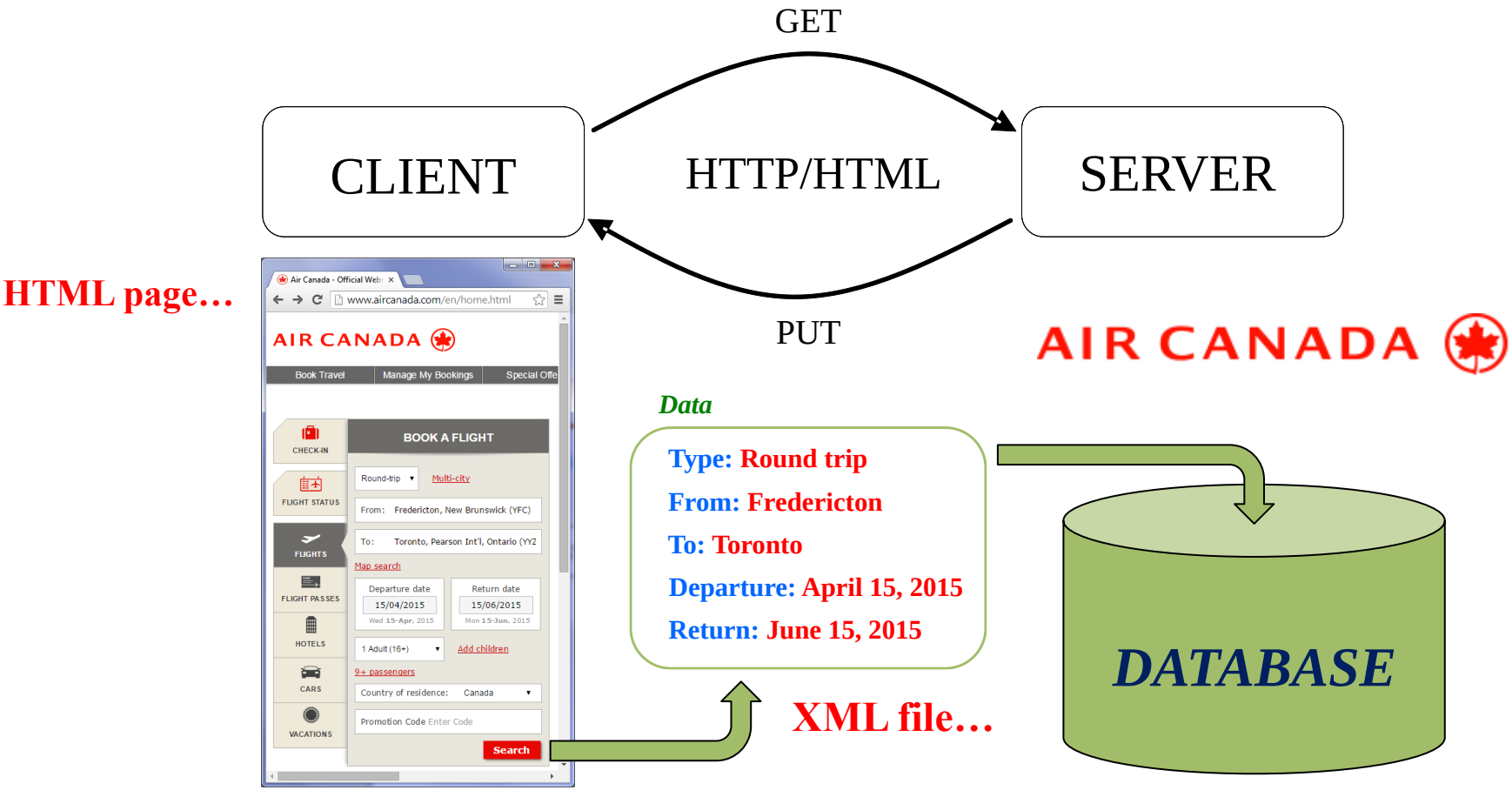
Introduction

- Air Canada → Online reservation



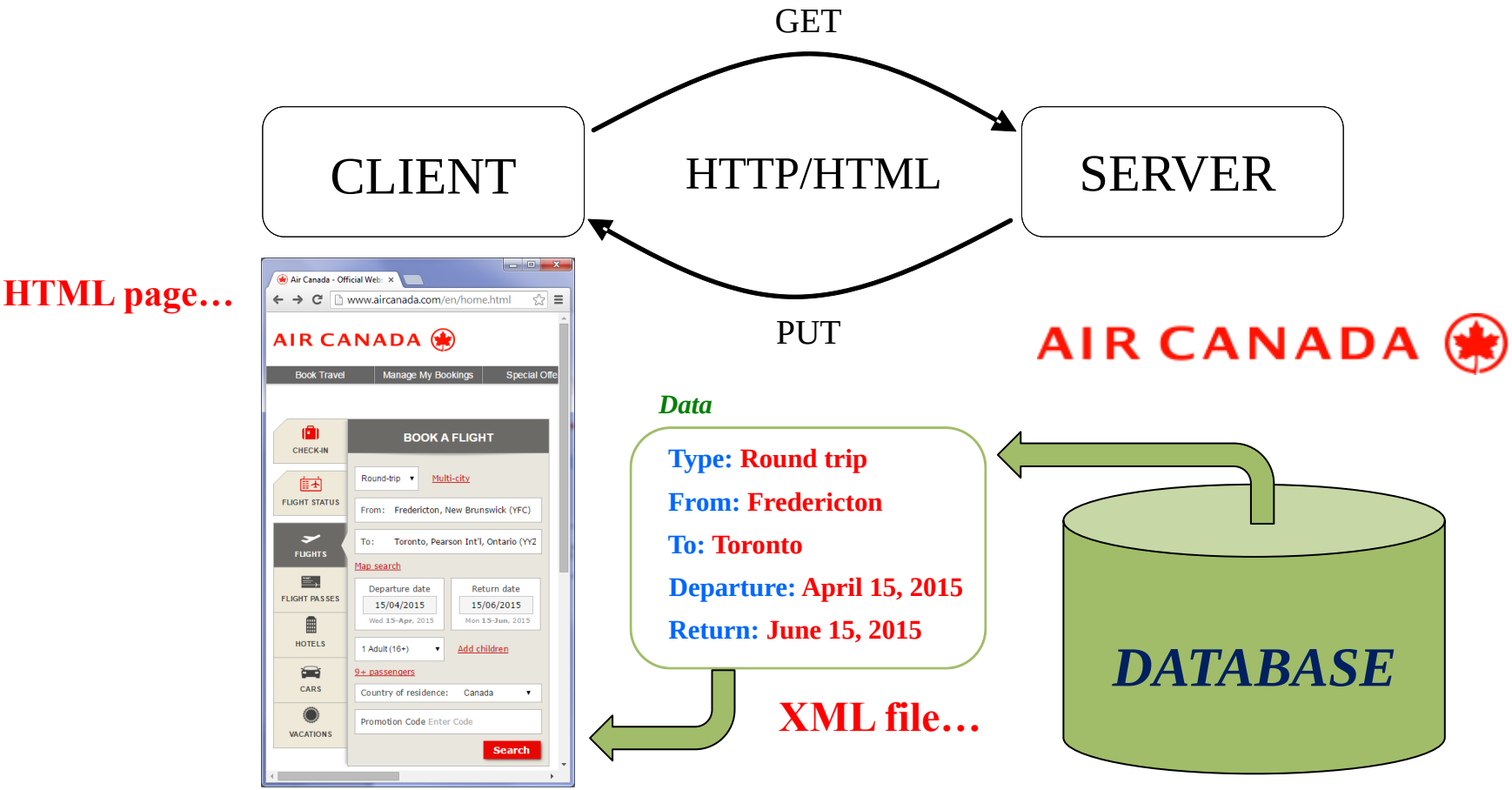
Introduction

- Air Canada → Online reservation



Introduction

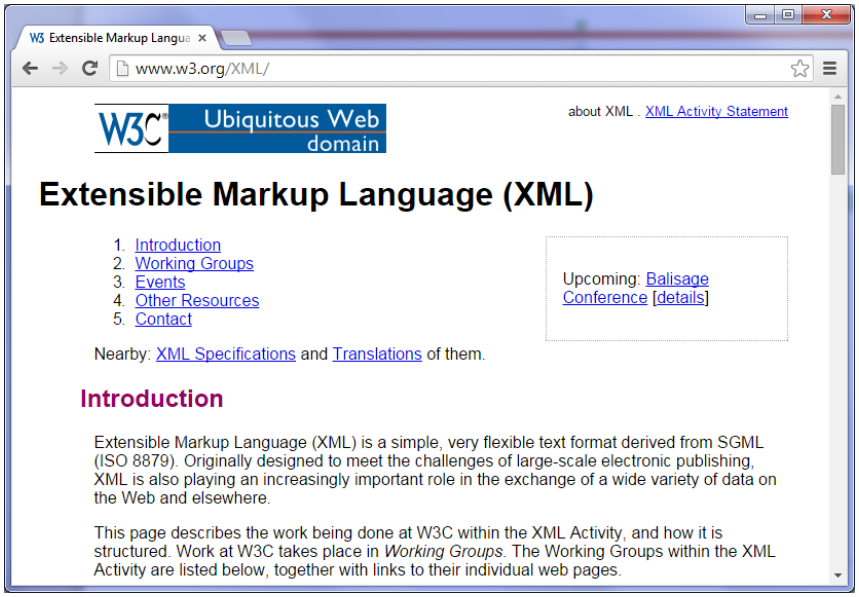
- Air Canada → **Review** my reservation



XML and Data

- XML…
  - eXtensible Markup Language
  - Developed by the World Wide Web Consortium (W3C)

- Nowadays…
  - XML is widely used for **describing** and **exchanging** data



http://www.w3.org/XML/

XML and Data

- What is so advantageous about XML…
  - It is portable
    - it utilizes unicode
  - It is platform independent
  - It is human readable
    - it is a pure and editable text
  - It is extensible
    - extra info can be added to a format without breaking applications based on previous versions
  - It is well supported
    - A large number of off-the-self tools for processing XML exist

XML and Data

- XML…
  - Has been built to support traditional applications (office and banking)
- What about applications involving **non-traditional** data ?
  - Other formats … based on XML have been proposed
  - E.g.,
    - Open GIS Consortium (OGC) recently published the Geography Markup Language (GML)

XML Infrastructure

- XML…
  - A W3C standard to complement HTML
  - XML is an application profile or restricted form of SGML
    - Standard Generalized Markup Language [ISO 8879]
  - A universal format for structured documents and data on the Web
- Motivation
  - **HTML** describes the **presentation**
  - **XML** describes the **content**

XML Infrastructure

- XML Syntax
  - XML is a textual representation of data
  - Basic component in XML is the **element**
    - Element is a piece of text bounded by matching tags, e.g., <author>Abiteboul</author>
    - Elements may be nested
    - Elements can be empty <value></value>, abbr. <value/>
    - An XML document is a single root element
  - Well formed XML doc: it has matching tags

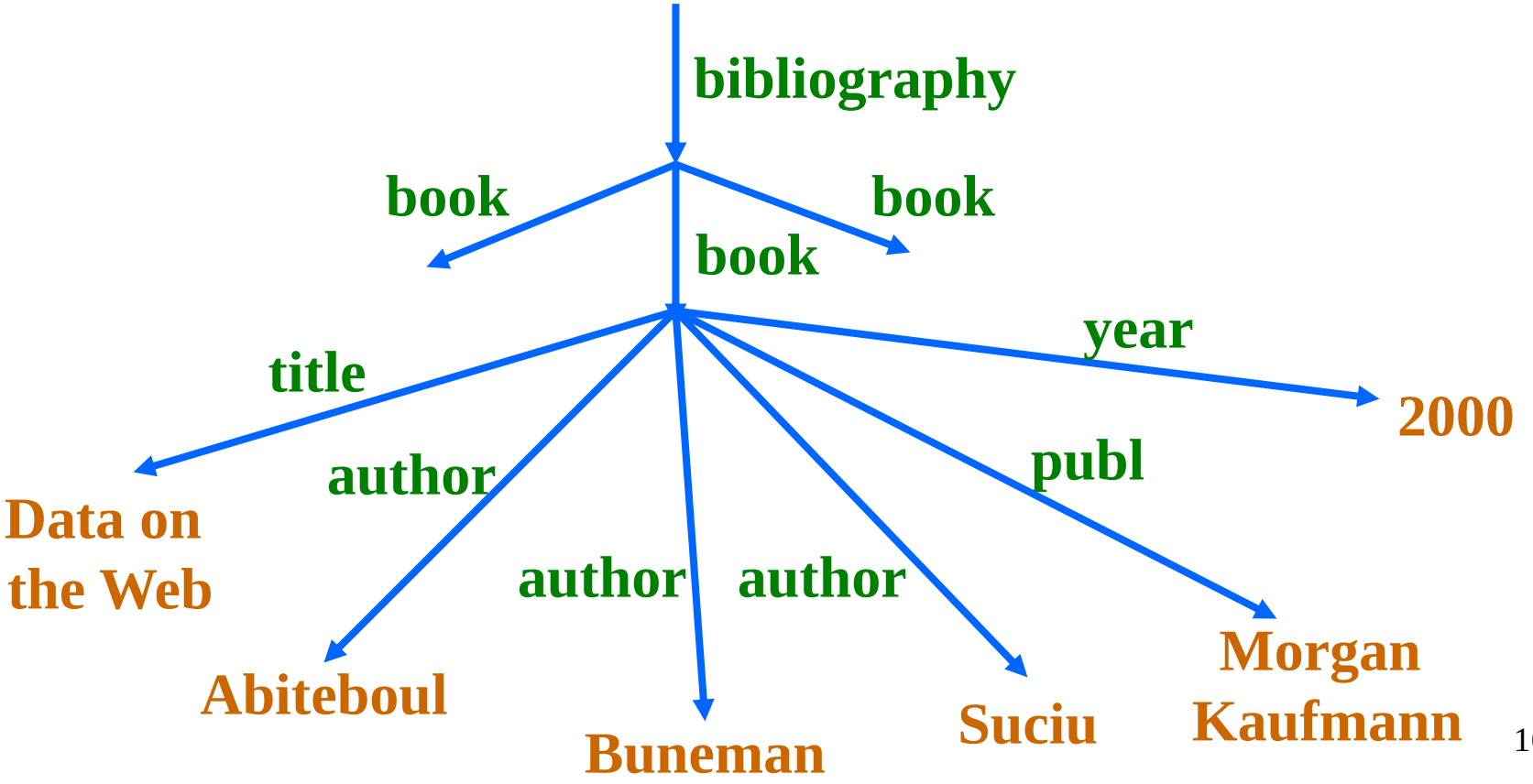
XML Infrastructure

- XML Syntax example (nested elements)

```
<bibliography>
   <book>
      <title>Data on the Web</title>
      <author>Abiteboul</author>
      <author>Buneman</author>
      <author>Suciu</author>
      <publ>Morgan Kaufmann</publ>
      <year>2000</year>
   </book>
   …
</bibliography>
```

XML Infrastructure

- XML diagram (**tree**) example …



XML Infrastructure

- XML Syntax
  - XML allows to associate **attributes** with elements, e.g.,

```
<book price="40" currency="Euro">
   <title>Data on the Web</title>
   <author>Abiteboul</author>
   …
</book>
```

  - Attributes are alternative ways to represent data

XML Infrastructure

- XML Syntax
  - XML allows to associate **unique identifiers** to elements as the value of a certain attribute
  - Using the attribute **idref** it is possible to **refer to** that element
  - This is an XML mechanism for describing

    Graphs rather than **trees**

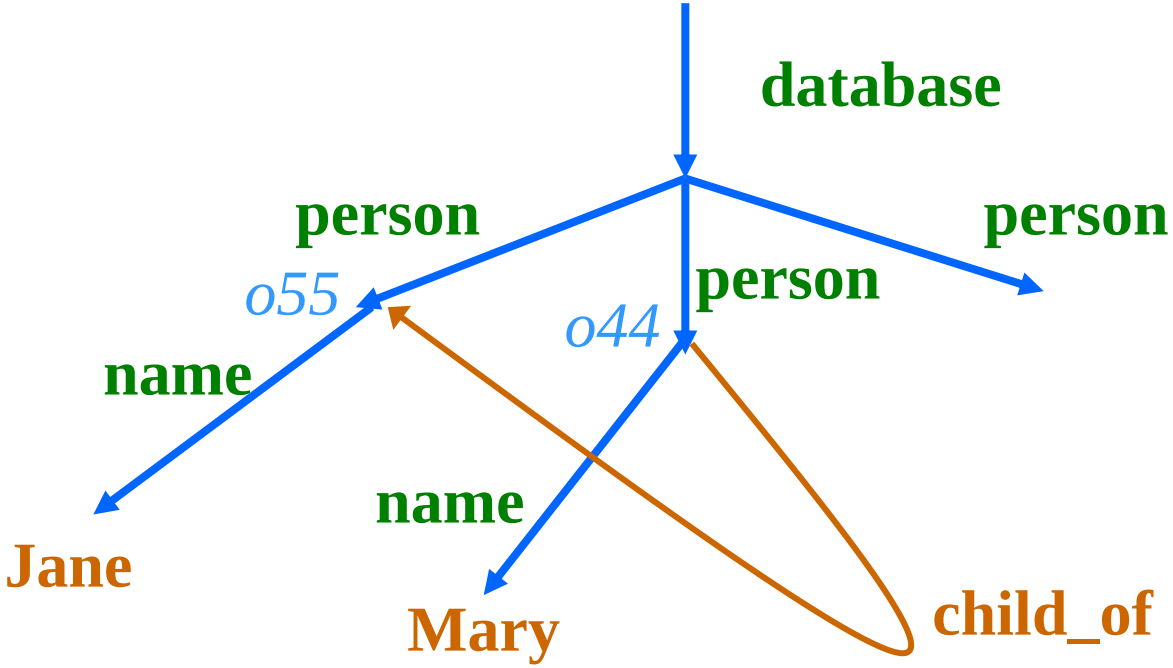
XML Infrastructure

- XML Syntax example

```
<database>
    <person id="o55">
        <name>Jane</name>
    </person>
    <person id="o44">
        <name>Mary</name>
        <child_of idref="o55"/>
    </person>
    …
</database>
```

XML Infrastructure

- XML diagram (**graph**) example …



XML Family of Technologies

- XML … is a **growing set of modules**
  - XML1.0
    - Is the specification that…
    - Defines what “tags” and “attributes” are
  - Xlink
    - Describes a standard way to add hyperlinks to an XML file
  - Xpointer and XFragments
    - Syntaxes in development for pointing to parts of an XML document
    - Xpointer is like a URL, which points to pieces of data inside an XML document

XML Family of Technologies

- XML Family Modules
  - CSS
    - The style sheet language
    - It is applicable to XML as it is to HTML
  - XSL
    - An advanced language for expressing style sheets
  - XSLT
    - A transformation language used for rearranging, adding, deleting tags and attributes
  - XML-QL
    - A powerful query language for info extraction from XML files

XML Family of Technologies

- XML Family Modules (cont')
  - DOM
    - A standard set of function calls for manipulating XML (and HTML) files from a programming language
  - XML Schemas 1 and 2
    - Help developers to precisely define the structures of their own XML formats
    - They provide a means for defining the structure, content and semantics of XML documents
  - … and many others… http://www.w3c.org

XML and Databases

- **Query 1: Is XML a Database ?**
  - … only in the strictest sense of the term…
    - An XML document is a collection of data
    - … like any other file…
  - As a "database" format…
    - XML has some **advantages**…
      - It is self-describing
        - » The markup describes the structure and type names of the data; although not the semantics
      - It is portable
        - » It utilizes unicode
      - It can describe data in tree or graph structures

XML and Databases

- Is XML a Database ?
  - As a “database” format…
    - XML also has some **disadvantages**…
      - Its elements are ordered…
      - It is verbose and has a peculiar syntax…
        - » XML can mix text and elements
        - » There is an ambiguity what to use… (attributes or elements?)
      - It has lots of other stuff…
        - » Entities, processing instructions, comments,…
      - The access to data is slow…
        - » … due to parsing and text conversions

XML and Databases

- XML elements are ordered…
  - The following XML docs are not equivalent

```
<person>
   <firstname>John</firstname>
   <lastname>Smith</lastname>
</person>

<person>
   <lastname>Smith</lastname>
   <firstname>John</firstname>
</person>
```

XML and Databases

- XML can mix text and elements…
  - several syntactic peculiarities from a DB perspective…

```
<talk>SSD and XML in Geography
   <speaker>John Smith</speaker>
</talk>
```

XML and Databases

- Use attributes or elements ?
  - …in order to represent information in XML

```
<person>
   <name>John</name>
   <age>33</age>
</person>
```

Or

```
<person name="John" age="33"/>
```

Or

```
<person age"33">
   <name>John</name>
</person>
```

XML and Databases

- **Query 2: Is XML a DBMS ?**

  [Does XML and its surrounding technologies… constitute a Database Management System (DBMS) ?]

  – … the answer is … "sort of"

  – The plus side…

    - **XML provides:** data storage, schemas, query languages, programming interfaces, …

  – The minus side…

    - **XML lacks of:** efficient storage, indexes, security, transactions and data integrity, multi-user access, triggers, queries across multiple documents, …

XML and Databases

- **Structuring XML**…
  - There are **two mechanisms** to constrain the contents (i.e., specify valid elements) in an XML document…
    - Document Type Definitions (DTD)
    - XML Schemas
  - An XML document …
    - that conforms to a DTD or an XML schema

  …is considered to be valid

XML and Databases

- **Structuring XML…**
  - Sample XML Fragment…

```
<parcel id= "P123x">
   <owner>John Smith</owner>
   <area>1200</area>
</parcel>
```

XML and Databases

```
<parcel id= "P123x">
    <owner>John Smith</owner>
    <area>1200</area>
</parcel>
```

- Structuring XML with…
  - **Document Type Definitions (DTD)**…
    - The original means of specifying the structure of an XML document
    - Used to specify the order and occurrence of elements in an XML document
    - It has a different syntax than XML

```
<!ELEMENT parcel (owner, area)>
<!ATTLIST parcel id CDATA>
<!ELEMENT owner (#PCDATA)>
<!ELEMENT area (#PCDATA)>
```

XML and Databases

- Structuring XML with…
  - **Document Type Definitions (DTD)**…
    - Proved to be **inadequate** for the needs of XML
    - Main reasons…
      - It has a different syntax than XML
      - It does not support data types
    - Microsoft Corporation submitted to W3C
      - A potential XML Schema standard
      - …named **XDR**
      - XDR tackled some of the problems of DTD
      - Finally, it was not accepted by W3C

XML and Databases

- Structuring XML with…
  - **XML Schema Definitions (XSD)**…
    - XML Schema is a W3C recommendation
    - XML Schema features…
      - It describes the structure and constraints on the content model of XML documents
      - It supports more data types that XDR
      - It allows the creation of custom data types
      - It supports object oriented concepts (like inheritance and polymorphism)

XML and Databases

```
<parcel id= "P123x">
    <owner>John Smith</owner>
    <area>1200</area>
</parcel>
```

- Structuring XML with…
  - **XML Schema Definitions (XSD)**…

```
<schema xmlns="http://www.w3.org/2001/XMLSchema">
   <element name="parcel">
      <complexType>
         <sequence>
            <element name="owner" type="string"/>
            <element name="area" type="unsignedInt"/>
         </sequence>
         <attribute name="id">
            <simpleType>
               <restriction base="string">
                  <pattern value="P\d{3}[A-Za-z]{1}"/>
               </restriction>
            </simpleType>
         </attribute>
      </complexType>
   </element>
</schema>
```

XML and Databases

- **Mapping XML Schemas to DB Schemas**
  - Mappings are performed on…
    - element types, attributes and text
    - (physical structure is omitted as well as some logical structure) … databases are concerned only with data
  - **Table-based mapping…**
    - XML documents are modeled …
      - … either as a single table
      - … or as a set of tables

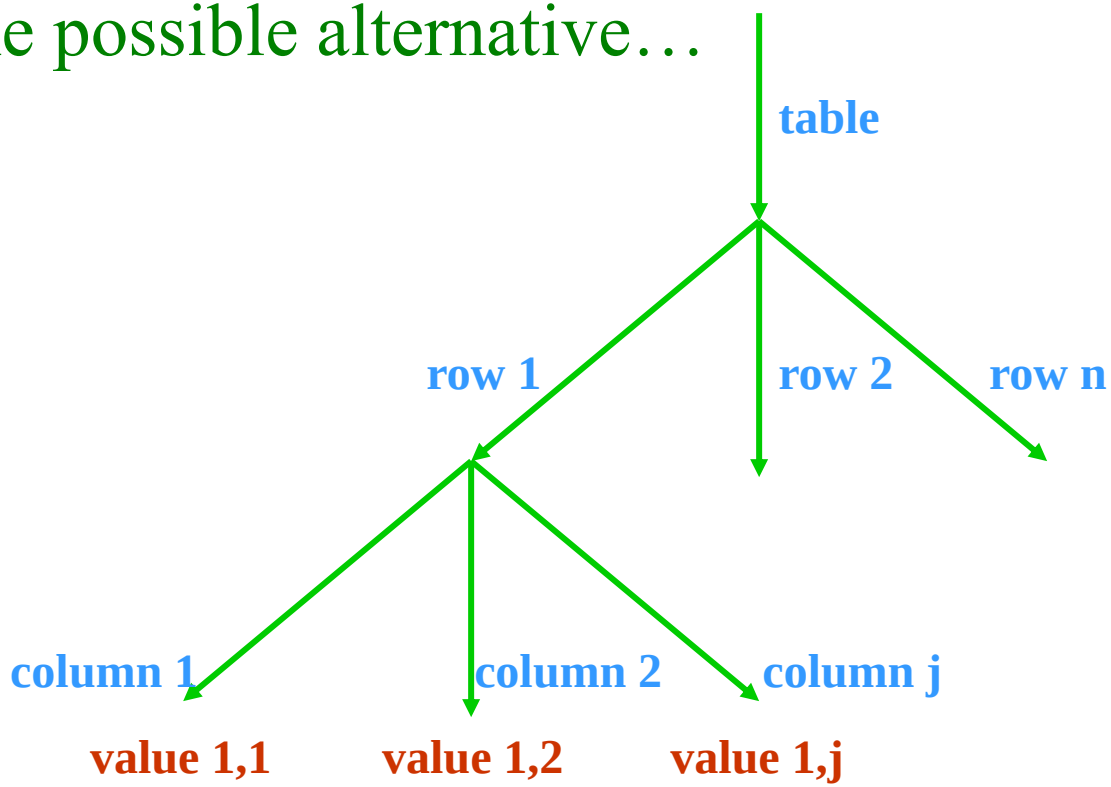
XML and Databases

- **Table-based mapping…**
  - One possible alternative…

```
<database>
    <table>
        <row>
            <column1>…</column1>
            <column2>…</column2>
                …
        </row>
        <row>
            …
        </row>
        …
    </table>
    …
</database>
```

XML and Databases

- **Table-based mapping…**
  - One possible alternative…



XML and Databases

- Representing Relation (POINTS) in XML

PID	X	Y
1	34	45
2	67	23
3	24	21

```
<table>
      <row>
            <PID>1</PID>
            <X>34</X>
            <Y>45</Y>
      </row>
      <row>
            <PID>2</PID>
            <X>67</X>
            <Y>23</Y>
      </row>
      <row>
            <PID>3</PID>
            <X>24</X>
            <Y>21</Y>
      </row>
</table>
```

XML and Databases

- **XML Querying**…
  - A number of languages have been developed …
    - XPath, XQuery, Lorel, UnQL, XML-QL, XQL, etc.

  …to extract information from XML documents

  - **XPath** (XML Path Language) …
    - A W3C recommendation
    - It utilizes a syntax that resembles hierarchical paths
    - Used to address parts of a file system or URL

XML and Databases

- XML Querying with…
  - **XPath** …
    - It provides functions (function library)
      - For interacting with selected data from a document
      - For accessing information about document nodes
      - For the manipulation of strings, numbers, and booleans
    - It is extensible with regards to functions
    - It uses a compact non-XML syntax
      - This facilitates the use of Xpath within URIs and XML attribute values (e.g., in XML Schema, XSLT)

XML and Databases

- XML Querying with…
  - **XPath** …
    - It operates on the abstract, logical structure of an XML document
    - It operates on a single XML document
    - It views the document as a tree of nodes
    - The values returned from an XPath query are considered as nodes
    - XPath data model considers many types of nodes
      - text nodes, element nodes, attribute nodes, root nodes, namespace nodes, processing instruction nodes, comment nodes

XML and Databases

- XML Querying with…
  - **XPath** …
    - Sample queries
      - Select all owner elements that are children of the root element parcel

        /parcel/owner
      - Select all owner elements

        //owner
      - Select all child elements of the root element parcel

        /parcel/*
      - Select all id attributes of the parcel elements in the document

        /parcel[@id]
      - Select all ancestors of all the owner elements that are children of the parcel element (which should select parcel element)

        /parcel/owner/ancestor::*

XML and Databases

- XML Querying …
  - A few words about **XSL**…
    - A W3C proposal …
      - Stylesheet specification language for XML
    - Its primary role…
      - Stylesheet transformations: XML→ HTML
      - General transformations: XML → XML
    - XSL data model…
      - Is an ordered tree…
      - Accurately corresponds to XML's
      - All XML constructs are addressed in XSL

XML and Databases

- XML Querying …
  - **XSL**…
    - An XSL program is…
      - A set of template rules
      - **Template rule = pattern + template**
    - XSL …
      - A recursive function…
    - Specifically,
      - XSL starts from the root element
      - It tries to apply a pattern to that node
      - If it succeeds, it executes the corresponding template
      - The template instructs XSL to produce an XML result
      - Recursive execution to node's (root's) children

XML and Databases

- XML Querying …
  - **XSL** example…
    - An XML document…

```
<cadastre>
        <parcel>
                <owner>John</owner>
                <use>residential</use>
        </parcel>
        <parcel>
                <owner>Mary</owner>
                <use>parking lot</use>
        </parcel>
        <parcel>
                <owner>Michael</owner>
                <use>agricultural</use>
        </parcel>
</cadastre>
```

XML and Databases

- XML Querying …
  - **XSL** example…
    - An XSL program
      - it returns owner names

```
<xsl:template>
        <xsl:apply-templates/>
</xsl:template>
<xsl:template match="/cadastre/*/owner">
        <result>
                <xsl:value-of/>
        </result>
</xsl:template>
```

References

Abiteboul, S., Buneman, P., and Suciu, D., 2000. *Data on the Web: From Relations to Semi-Structured Data and XML*. Morgan-Kaufmann.

Bourett, R., 2001. XML and Databases. http://www.rpbourret.com/xml/XMLAndDatabases.htm

Bourett, R., 2001. XML Database Products. http://www.rpbourret.com/xml/XMLDatabaseProds.htm

Obasanjo, D., 2001. An Exploration of XML in Database Management Systems. http://www.25hoursaday.com/StoringAndQueryingXML.html

Stefanakis, E., 2002. Tutorial: Semi-structured Data and XML in Geographic Data Modeling and Handling. *Join International Symposium on Geospatial Theory, Processing and Applications*, Ottawa, Canada. http://www.dbnet.ece.ntua.gr/ ~stefanak/TU1_Stefanakis.htm

Suciu, D., 2001. On Database Theory and XML. *SIGMOD Record*. 30(3): 39-45.

World Wide Web Consortium (W3C), http://www.w3c.org/