Unlock Content™

**M A R K**
LOGIC

# XML and XQuery: Unlocking the Value of Content

# Agenda

- ■ **Welcome and introduction to Mark Logic**

- ■ XML and XQuery: unlocking the value of content

  - □ Stephen Buxton, director of product management for Mark Logic

  - □ Co-author, "Querying XML: XQuery, XPath, and SQL/XML in Context"

- ■ Q&A

# A Coming Generation of Content Apps



Business managers are overwhelmed by the sheer volume of information available to them, as they are forced to search for, cut/paste, or recreate content that could have been served to them *in context* .

# Content is Hard

*Everything about content is hard: its hard to integrate, hard to find, hard to repurpose, and hard to deliver*

# It's All About the Infrastructure

| | **Data** | **Content** |
|---|---|---|
| **Data model** | Relational | XML |
| **Access method** | SQL | XQuery |
| **Platform** | Oracle, DB2, SQL Server, etc. | MarkLogic Server, ??? |

# About Mark Logic

**Mark Logic provides of the industry's leading *XML content server***

- 30+ top-tier customers: *DISA, U.S. Air Force, DTIC, U.S. Army, Elsevier, JetBlue, Oxford University Press*

- 45+ employees

- Headquarters in San Mateo, California (Silicon Valley)

- Strong financial backing:  *Sequoia Capital, Lehman Brothers*

# Recognized Technology Leader

- Named to Gartner's list of "Cool Vendors"

- Named a "Company to Watch" by Intelligent Enterprise magazine

- Named to KMWorld magazine's "100 Companies that Matter in Knowledge Management"

- Won 2005 Software and Information Industry Association Codie Award for Best Vertical Market Product or Service  - Publishing

- Named a "Trend-Setting Product" by KMWorld magazine

- Named to prestigious list of companies that matter most in the digital content industry

# Selected Mark Logic Customers

## "Big, Fast, XML"

- ***Big*** – from gigabytes to terabytes … and beyond

- ***Fast*** – capable of millisecond response times … leveraging search-engine indexing techniques

- ***XML*** – designed specifically for XML content … implementing the W3C-standard XQuery language

# How We Help Our Customers

**Mark Logic accelerates the creation and delivery of the specific information needed to accomplish vital missions**

- Integrate content from diverse sources

- Repurpose content for multiple applications

- Enable dynamic custom views of content

- Deliver information through any channel or device

- Search-and-discover previously unknown information

# Why Choose MarkLogic Server?

**Only MarkLogic Server delivers**

- ■ "As-is" content loading
  - ☐ Does not require DTDs or schemas
  - ☐ No need to "shred" or "chunk"
  - ☐ Automatic conversion to XML
- ■ Industry's most powerful and complete XQuery engine
  - ☐ W3C standard for XML query
  - ☐ Flexible programming language designed for content
  - ☐ More powerful queries, more precise results
- ■ High scalability and performance
  - ☐ Millisecond response times – multi-terabyte contentbases
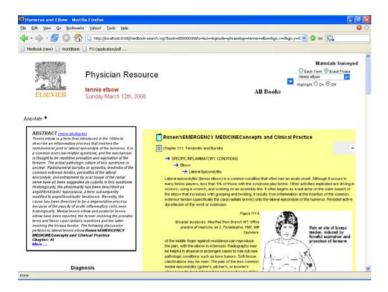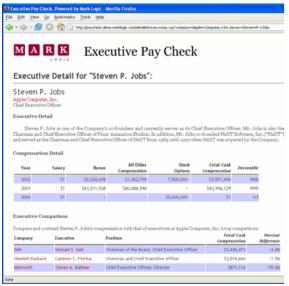
# Agenda

- Welcome and introduction to Mark Logic

- XML and XQuery: unlocking the value of content

  - Stephen Buxton, director of product management for Mark Logic

  - Co-author, "Querying XML: XQuery, XPath, and SQL/XML in Context"
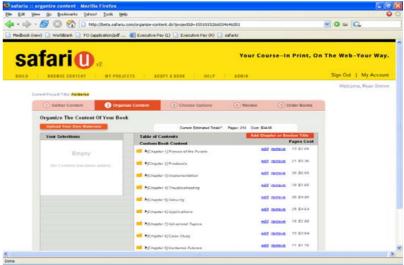
- Q&A

# Content Applications

# More, richer content



<name>Smith</name> came in from <place>Kansas</place> on <date>3rd June 2001</date> by plane

3 years

3 months

# "locked-in" value

database

\+ search engine

_____

\=  or  ??

XML

+ XQuery

+ XML content server

————————————

platform for content applications

# Demos

Medbook

World Bank

Message Traffic Analysis

XML

\+         XQuery

\+    XML content server

––––––––––––––––––––––––

platform for content applications

XML

+      XQuery

+      XML content server
_____

platform for content applications

# XML data, XML content

```xml
<?xml version="1.0"?>

<ipo:purchaseOrder xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:ipo="http://www.example.com/IPO" orderDate="1999-12-01">

  <shipTo exportCode="1" xsi:type="ipo:UKAddress">
    <name>Helen Zoe</name>
    <street>47 Eden Street</street>
    <city>Cambridge</city>
    <postcode>CB1 1JR</postcode>
  </shipTo>

  <billTo xsi:type="ipo:USAddress">
    <name>Robert Smith</name>
    <street>8 Oak Avenue</street>
    <city>Old Town</city>
    <state>PA</state>
    <zip>95819</zip>
  </billTo>

  <items>
    <item partNum="833-AA">
      <productName>Lapis necklace</productName>
      <quantity>1</quantity>
      <USPrice>99.95</USPrice>
      <ipo:shipComment>
        Use gold wrap if possible
      </ipo:shipComment>
      <ipo:customerComment>
        Want this for the holidays!
      </ipo:customerComment>
      <shipDate>1999-12-05</shipDate>
    </item>
  </items>

</ipo:purchaseOrder>
```

# XML data, XML content

```xml
<!DOCTYPE book PUBLIC "-//OASIS//DTD DocBook XML V4.1.2//EN" "http://www.oasis-open.org/docbook/xml/4.1.2/docbookx.dtd">

<book>

  <bookinfo>

    <title>DocBook document example</title>
    <author>
      <firstname>Ovidiu</firstname>
      <surname>Predescu</surname>
    </author>
    <copyright>
      <year>2001</year>
      <holder>Ovidiu Predescu</holder>
    </copyright>

  </bookinfo>

  <preface>

    <title>Example document</title>
    <para>A simple DocBook example document.</para>

  </preface>

  <chapter>

    <title>DocBook and XSLT-process mode for GNU Emacs/XEmacs</title>

    <para>The <emphasis>XSLT-process</emphasis> mode, when used with the DocBook-XSL stylesheets makes GNU Emacs/XEmacs
the ideal DocBook processor.</para>

    <para>To find out more on how to use the <emphasis>XSLT-process</emphasis> mode for DocBook processing, point your browser
to <ulink url="http://xslt-process.sourceforge.net/docbook.php">http://xsltprocess.sourceforge.net/docbook.php</ulink>.</para>

  </chapter>

</book>
```

# XML data, XML content

- Constraints

- Shredding

- Richness

# XML content gets better and better

**??**

community participation

user annotation

analytical markup

semantic markup

structured markup

metadata

plain text

# The XML eco-system



*Authoring*

*Visualization/PDF*

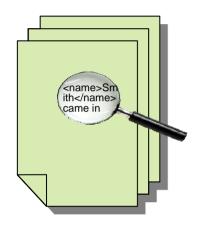*Enrichment*

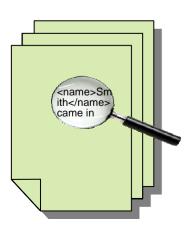*Services*

# XML
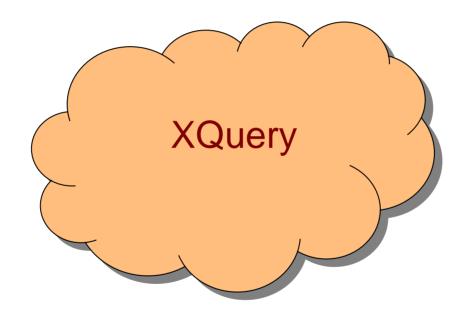# + XQuery
# + XML content server
---
# platform for content applications

# Query, manipulate, and transform XML with XQuery
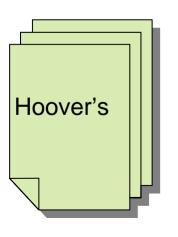
XQuery

XML

XHTML

RSS

Web Service

XSL:FO to PDF

Mobile

# Query, manipulate, and transform XML with XQuery

**SEC filings**

**Hoover's**

**Convert:**
- Get DEF 14A filings from SEC
- Get executive, competitor listings from Hoover's
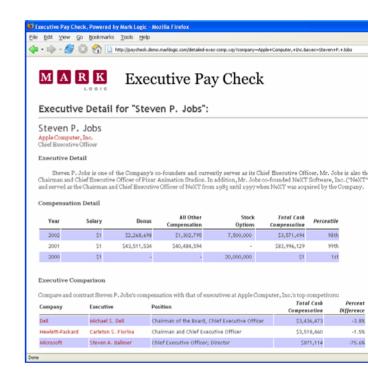- Load, automatically converting HTML → XML

**Query:**
1. Identify and extract position title
2. Identify and extract paragraphs about Jobs
3. Identify and extract compensation table
8. Use Hoover's to identify top 3 competitors
9. Identify DEF-14A documents for competitors
10. Identify comparable executives
11. Extract their compensation information

**Manipulate:**
4. Identify table rows concerning Steven Jobs
5. Correlate table columns to generic standard
6. Calculate total cash compensation
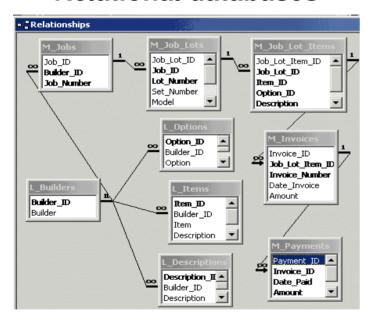7. Compare compensation to aggregate
12. Calculate comparative pay

**Render:**
13. Generate automatic cross-linkages
14. Generate output as HTML

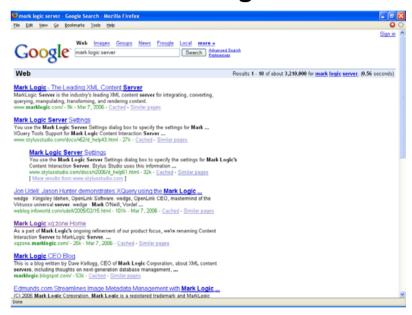Executive Pay Check, Powered by Mark Logic - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://paycheck.demo.marklogic.com/detailed-exec-comp.xqy?company=Apple+Computer,+Inc.&exec=Steven+P.+Jobs

MARK LOGIC   **Executive Pay Check**

**Executive Detail for "Steven P. Jobs":**

Steven P. Jobs
Apple Computer, Inc.
Chief Executive Officer

**Executive Detail**

Steven P. Jobs is one of the Company's co-founders and currently serves as its Chief Executive Officer. Mr. Jobs is also the Chairman and Chief Executive Officer of Pixar Animation Studios. In addition, Mr. Jobs co-founded NeXT Software, Inc. ("NeXT") and served as the Chairman and Chief Executive Officer of NeXT from 1985 until 1997 when NeXT was acquired by the Company.

**Compensation Detail**

| Year | Salary | Bonus | All Other Compensation | Stock Options | Total Cash Compensation | Percentile |
|------|--------|-------|------------------------|---------------|-------------------------|------------|
| 2002 | $1 | $2,268,698 | $1,302,795 | 7,500,000 | $3,571,494 | 98th |
| 2001 | $1 | $43,511,534 | $40,484,594 | - | $83,996,129 | 99th |
| 2000 | $1 | - | - | 20,000,000 | $1 | 1st |

**Executive Comparison**

Compare and contrast Steven P. Jobs's compensation with that of executives at Apple Computer, Inc.'s top competitors:

| Company | Executive | Position | Total Cash Compensation | Percent Difference |
|---------|-----------|----------|-------------------------|--------------------|
| Dell | Michael S. Dell | Chairman of the Board, Chief Executive Officer | $3,436,473 | -3.8% |
| Hewlett-Packard | Carleton S. Fiorina | Chairman and Chief Executive Officer | $3,518,460 | -1.5% |
| Microsoft | Steven A. Ballmer | Chief Executive Officer; Director | $871,114 | -75.6% |

Done

# Databases, search engines, combinations fall short

## Relational databases



*Built for data*

## Search engines



*Built for documents*

# Let's get specific

More than simple "XML-aware" full-text query:

- XML structure and phrases

- XML structure and keywords

- Fine-grained search and retrieval

# Need more than "XML-aware" keyword search

*Show me sections that say "retract the landing gear"*

XML indicates a phrase break:

```
<step number="3">
  Wait for the automatic pilot control to
  retract.
</step>
<step number="4">
  The landing gear should now be ready
  to deploy.
</step>
```

Phrase should not match, because of change in context:

```
<warning>
  While waiting for the flaps to fully
  retract<footnote>The landing gear need
  to be stowed or else the flaps will
  remain deployed</footnote>, ensure
  that the automatic pilot control is at
  hand.
</warning>
```

XML does not indicate a phrase break:

```
<p>
  It is vital that you retract <b>the landing
  gear</b>, but leave the flaps fully
  deployed, before attempting this type
  of approach.
</p>
```
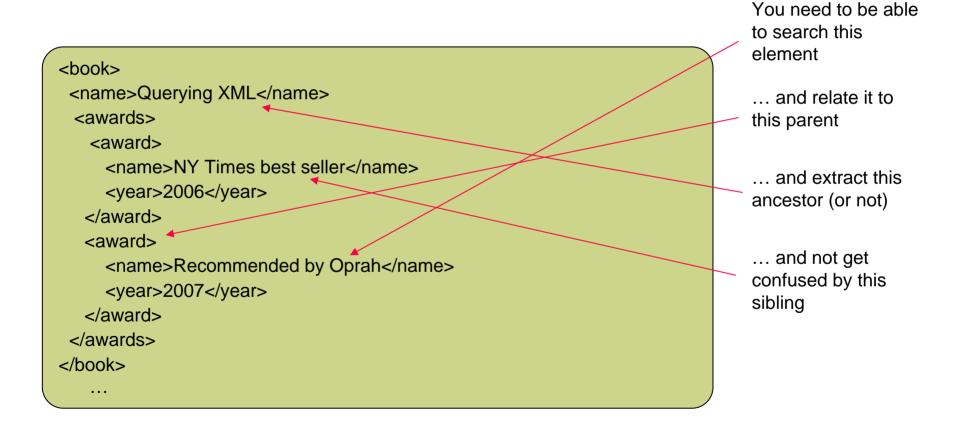
Phrase should match *around* the change in context:

```
<step>
  Partially retract<footnote>It is not
  recommended to fully retract at this
  point.</footnote> the landing gear.
</step>
```
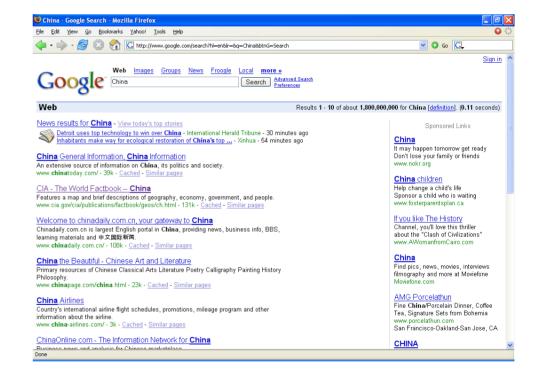
# Need more than "fielded search"

*Show me titles of books recommended by Oprah in 2006*

```
<book>
 <name>Querying XML</name>
  <awards>
   <award>
     <name>NY Times best seller</name>
     <year>2006</year>
   </award>
   <award>
     <name>Recommended by Oprah</name>
     <year>2007</year>
   </award>
  </awards>
</book>
    …
```

You need to be able to search this element

… and relate it to this parent

… and extract this ancestor (or not)

… and not get confused by this sibling

# Arbitrary granularity

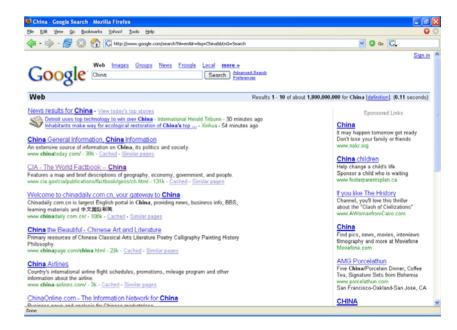*Show me all the documents that contain the word "China"*

# Fine-grained query

*Show me all the documents that contain the word "China" in the title.*

- Pre-process as metadata (search engine or database)

- "XML-aware full-text index"

# Fine-grained query *and* retrieval

**Show me all the sections that contain the word "China".**

```
<book>
 <title>Asian History</title>
 <section title="Introduction">
This book explains the history
of the …
 </section>
 <section title="Middle Ages">
For centuries China stood as a
leading …
 </section>
 <section title="Modern Day">
Today's China is …
 </section>
</book>
```

```
<journal>
 <title>Business Today</title>
 <section title="Background">
The world economy is …
 </section>
 <section title="Corporations">
Most American corporations
are still …
 </section>
 <section title="Global">
The biggest companies in
China are …
 </section>
</journal>
```

```
<news>
 <title>Baseball Update</title>
 <author>Tom China</author>
 <section>Today, the Red Sox
won their first ….
 </section>
 <section>Baseball to 'review'
Bonds allegations …
 </section>
 <section>Canada stuns U.S.
with World Baseball challenge
match win …
 </section>
</news>
```

# Document-level query

**Show me all the sections that contain the word "China".**

```
<book>
 <title>Asian History</title>
 <section title="Introduction">
This book explains the history
of the …
 </section>
 <section title="Middle Ages">
For centuries China stood as a
leading …
 </section>
 <section title="Modern Day">
Today's China is …
 </section>
</book>
```

```
<journal>
 <title>Business Today</title>
 <section title="Background">
The world economy is …
 </section>
 <section title="Corporations">
Most American corporations
are still …
 </section>
 <section title="Global">
The biggest companies in
China are …
 </section>
</journal>
```

```
<news>
 <title>Baseball Update</title>
 <author>Tom China</author>
 <section>Today, the Red Sox
won their first ….
 </section>
 <section>Baseball to 'review'
Bonds allegations …
 </section>
 <section>Canada stuns U.S.
with World Baseball challenge
match win …
 </section>
</news>
```

**Classical search engine solution.**

# Fine-grained query

*Show me all the sections that contain the word "China".*

```
<book>
 <title>Asian History</title>
 <section title="Introduction">
This book explains the history
of the …
 </section>
 <section title="Middle Ages">
For centuries China stood as a
leading …
  </section>
 <section title="Modern Day">
Today's China is …
  </section>
</book>
```

```
<journal>
 <title>Business Today</title>
 <section title="Background">
The world economy is …
 </section>
 <section title="Corporations">
Most American corporations
are still …
  </section>
 <section title="Global">
The biggest companies in
China are …
  </section>
</journal>
```

```
<news>
 <title>Baseball Update</title>
 <author>Tom China</author>
 <section>Today, the Red Sox
won their first ….
 </section>
 <section>Baseball to 'review'
Bonds allegations …
 </section>
 <section>Canada stuns U.S.
with World Baseball challenge
match win …
 </section>
</news>
```

**Database or search engine with "XML-aware" full-text.**

# Fine-grained query and retrieval, separate

*Show me all the sections that contain the word "China".*

```
<book>
 <title>Asian History</title>
<section title="Introduction">
This book explains the history
of the …
 </section>
<section title="Middle Ages">
For centuries China stood as a
leading …
 </section>
<section title="Modern Day">
Today's China is …
 </section>
</book>
```

```
<journal>
 <title>Business Today</title>
<section title="Background">
The world economy is …
 </section>
<section title="Corporations">
Most American corporations
are still …
 </section>
<section title="Global">
The biggest companies in
China are …
 </section>
</journal>
```

```
<news>
 <title>Baseball Update</title>
 <author>Tom China</author>
 <section>Today, the Red Sox
won their first ….
 </section>
 <section>Baseball to 'review'
Bonds allegations …
 </section>
 <section>Canada stuns U.S.
with World Baseball challenge
match win …
 </section>
</news>
```

**Database with XQuery + separate full-text.**

# Fine-grained query and retrieval, integrated

***Show me all the sections that contain the word "China".***

```
<book>
  <title>Asian History</title>
  <section title="Introduction">
This book explains the history
of the …
  </section>
<section title="Middle Ages">
For centuries China stood as a
leading …
  </section>
 <section title="Modern Day">
Today's China is …
  </section>
</book>
```

```
<journal>
  <title>Business Today</title>
  <section title="Background">
The world economy is …
  </section>
 <section title="Corporations">
Most American corporations
are still …
  </section>
<section title="Global">
The biggest companies in
China are …
  </section>
</journal>
```
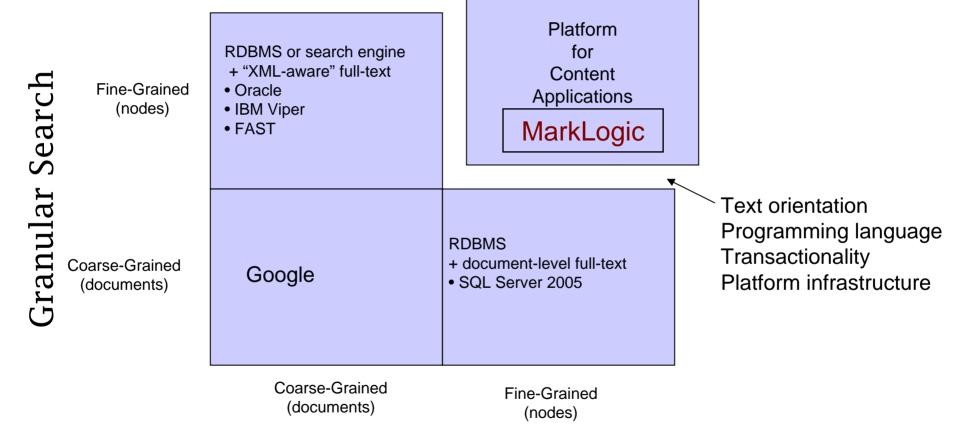
```
<news>
  <title>Baseball Update</title>
  <author>Tom China</author>
  <section>Today, the Red Sox
won their first ….
  </section>
  <section>Baseball to 'review'
Bonds allegations …
  </section>
  <section>Canada stuns U.S.
with World Baseball challenge
match win …
  </section>
</news>
```

**Contentbase with XQuery + integrated full-text.**

# Combinations



Granular Search

Fine-Grained
(nodes)

RDBMS or search engine
+ "XML-aware" full-text
• Oracle
• IBM Viper
• FAST

Platform
for
Content
Applications

MarkLogic

Coarse-Grained
(documents)

Google

RDBMS
+ document-level full-text
• SQL Server 2005

Text orientation
Programming language
Transactionality
Platform infrastructure

Coarse-Grained
(documents)

Fine-Grained
(nodes)

Granular Retrieval

XML

\+ XQuery

\+ XML content server

───────────────────
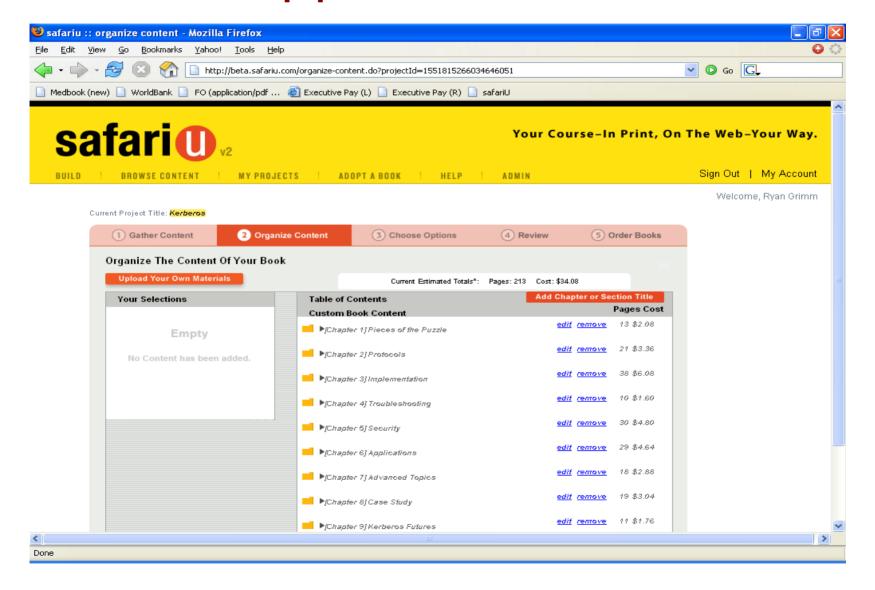
platform for content applications

# MarkLogic is architected for Big, Fast, XML content

# Content applications in the wild

# Agenda

- Welcome and introduction to Mark Logic

- XML and XQuery: unlocking the value of content

  - Stephen Buxton, director of product management for Mark Logic

  - Co-author, "Querying XML: XQuery, XPath, and SQL/XML in Context"

- Q&A

# Thank You

www.marklogic.com