To structure or not to structure, is that the question?

Paolo Atzeni

ROMA TRE
UNIVERSITÀ DEGLI STUDI

Based on work done with (or by!)
G. Mecca, P.Merialdo, P. Papotti, and many others

Lyon, 24 August 2009

Content

- A personal (and questionable …) perspective on the role structure (and its discovery) has in the Web, especially as a replacement for semantics
- Two observations
  - Especially important if we want automatic processing
  - Something similar happened for databases many years ago
- and a disclaimer:
  - I feel honored to have been invited, but I am not sure why it happened …
  - Personal means that I will refer to projects in my group (with some contradiction, as I have been personally involved only in some of them…)

Content

- Questions on the Web
- Structures in the Web, transformations: Araneus
- Information extraction, wrappers: Minerva, RoadRunner
- Crawling, extraction and integration: Flint

Web and Databases (1)

- The Web is browsed and searched
  - Usually one page or one item at the time (may be a ranked list)
- Databases are (mainly) queried
  - We are often interested in tables as results

What do we do with the Web?

- Everything!
- We often search for specific items, and search engines are ok
- In other cases we need more complex things:
  - What could be the reason for which Frederic Joliot-Curie (a French physicist and Nobel laureate, son in law of P. and M. Curie) is popular in Bulgaria (for example, the International House of Scientists in Varna is named after him)?
  - Find a list of qualified database researchers who have never been in the VLDB PC and deserve to belong to the next one

- I will not provide solutions to these problems, but give hints
  - Semantics
  - Structure

Questions on the Web: classification criteria (MOSES European project, FP5)

- Format of the response
- Number (and relationship) of the involved sites (one or more; of the same organization or not)
- Structure of the source information
- Extraction process

Format of the response

- A simple piece of data
- A web page (or fragment of it)
- A tuple of data
- A "concept"
- A set or list of one of the above
- Natural language (i.e., an explanation)

Involved sites

- One or more
- Of the same organization or not

- Sometimes also sites and local databases

Structure of the source information

- Information directly provided by the site(s) as a whole
- Information available in the sites and correlated but not aggregated
- Information available in the site in a unrelated way

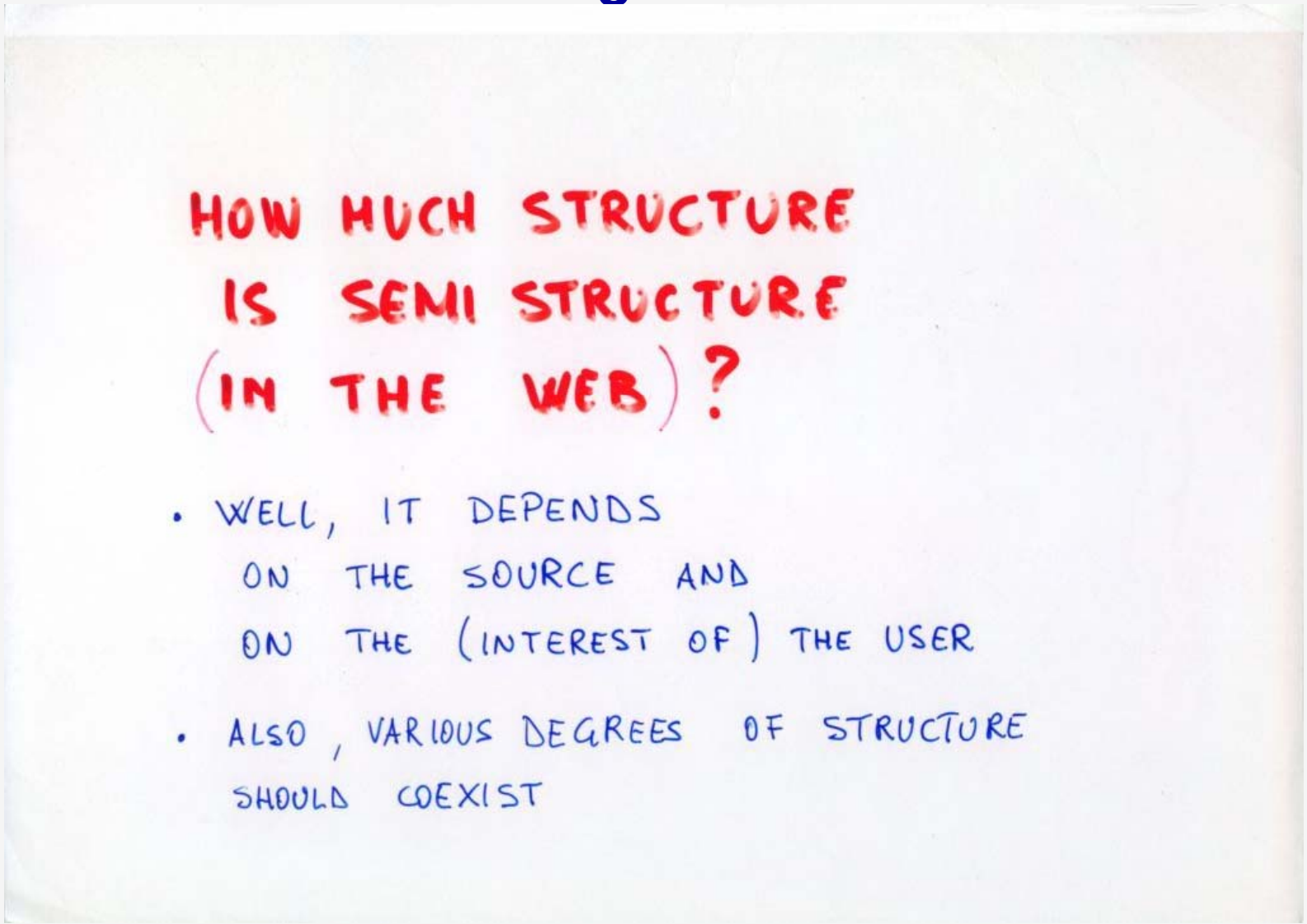
Extraction process

- Explicit service: this the case for example if the site offers an address book or a search facility for finding professors
- One-way navigation: here the user has to navigate, but without the need for going back and forth (no “trial and error”)
- Navigation with backtracking

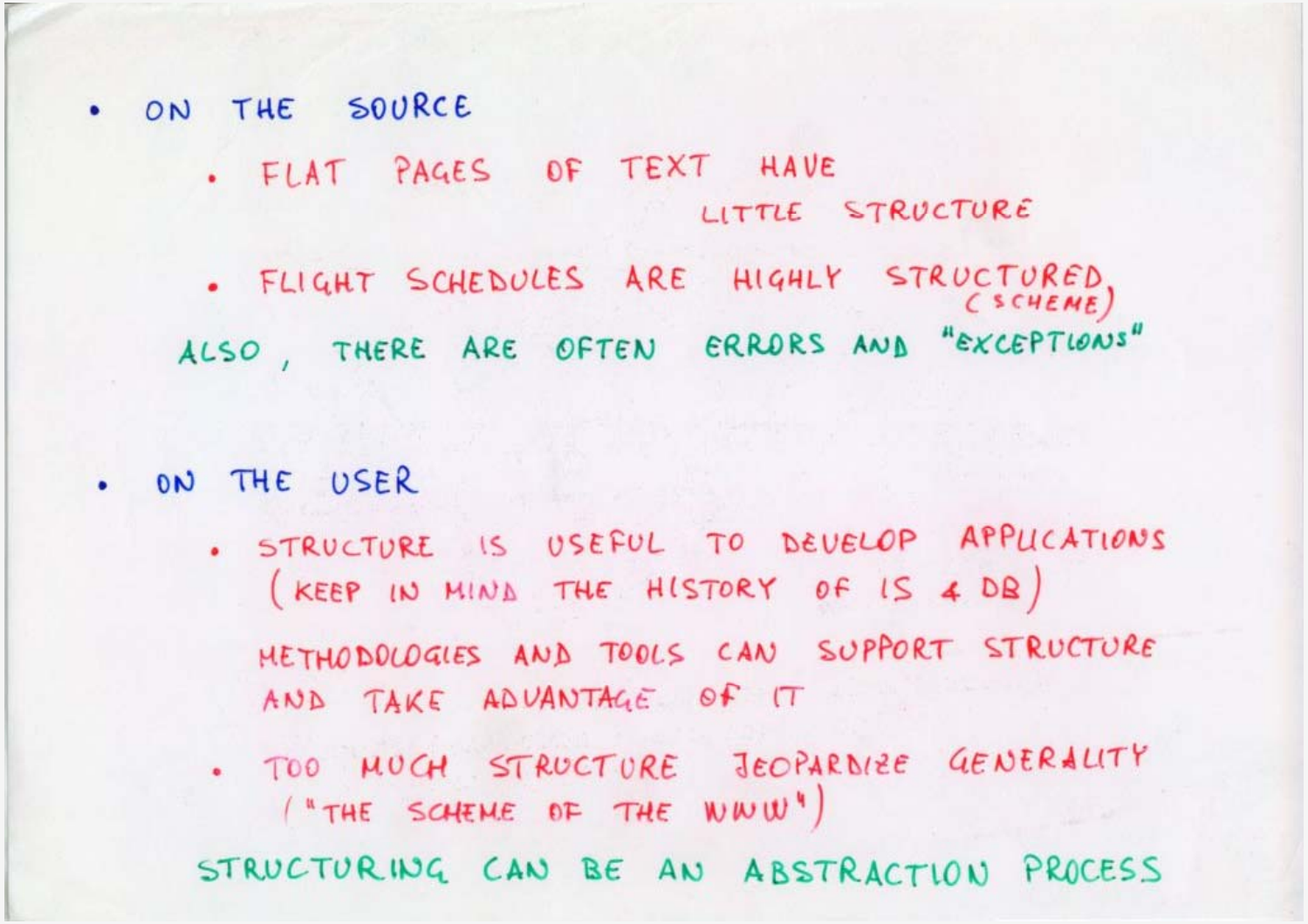
Web and Databases (2)

- The answer to some questions requires structure (we need to "compile" tables)
  - databases are structured and organized
  - how much structure and organization is there in the Web?

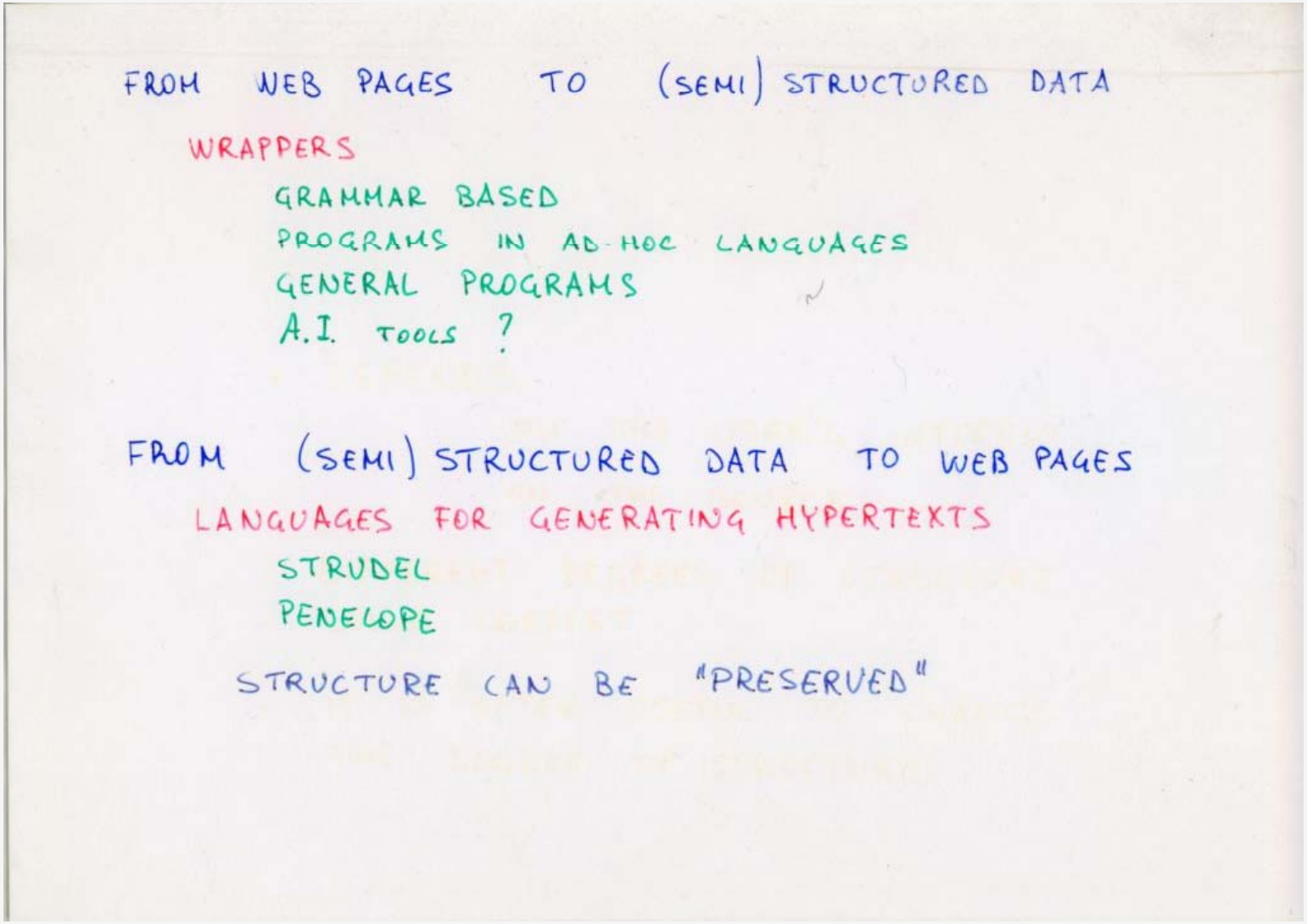
Workshop on Semistructured Data at Sigmod 1997



Workshop on Semistructured Data at Sigmod 1997 - 2



Workshop on Semistructured Data at Sigmod 1997 - 3



Transformations

- bottom-up: accessing information from Web sources
- top-down: designing and maintaining Web sites
- global: integrating existing sites and offering the information through new ones

The Araneus Project

- At Università Roma Tre & Università della Basilicata 1996-2003
- Goals: extend database-like techniques to Web data management
- Several prototype systems

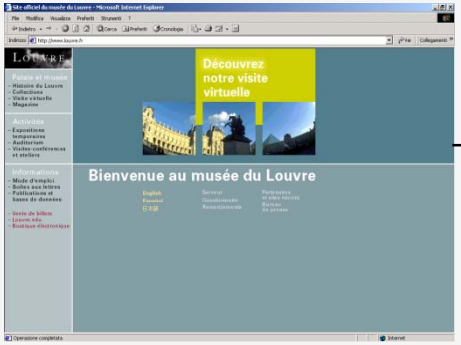
Example Applications

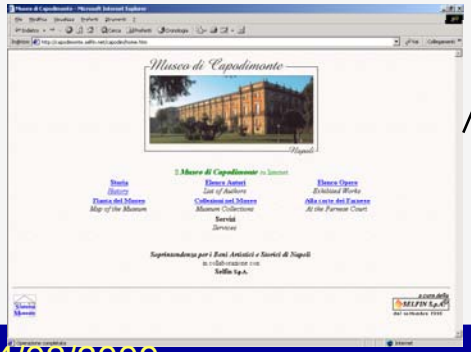
- The Integrated Web Museum
  a site integrating data coming from the Uffizi, Louvre and Capodimonte Web sites
- The ACM Sigmod Record XML Version
  XML+XSL Version of an existing (and currently running) Web site

The Integrated Web Museum

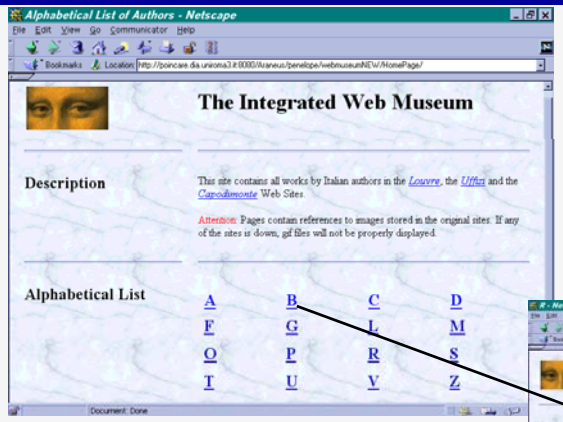
- A new site with data coming from the Web sites of various museums
  - Uffizi, Louvre and Capodimonte

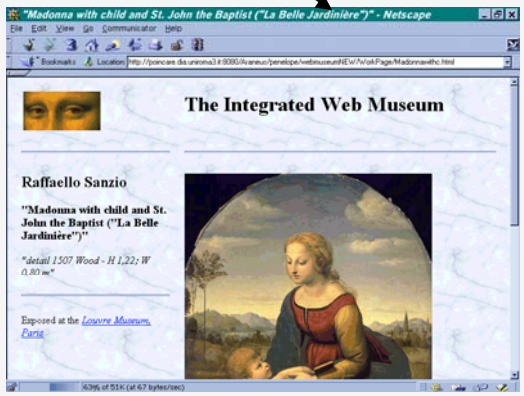
The Integrated Web Museum





DB



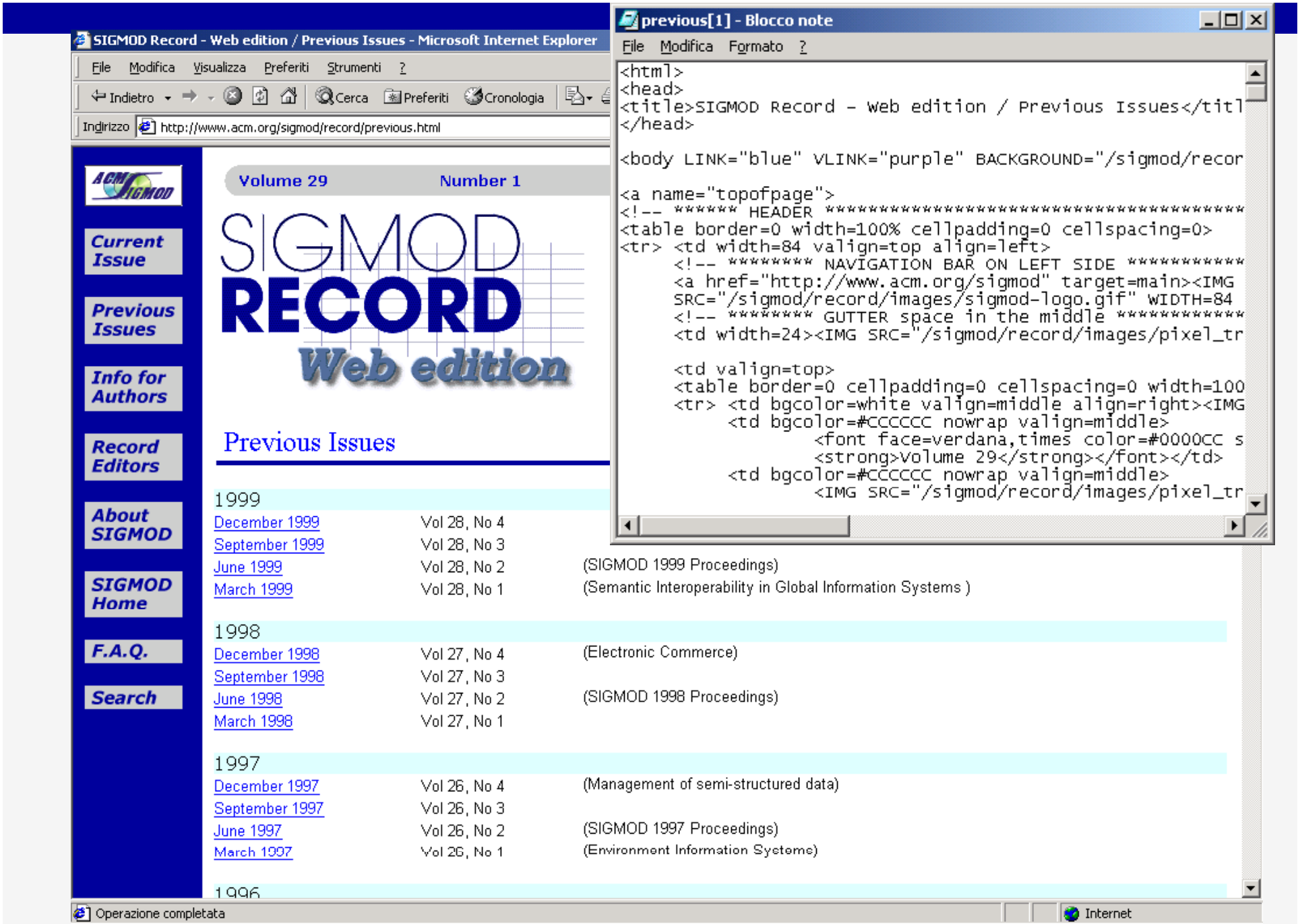


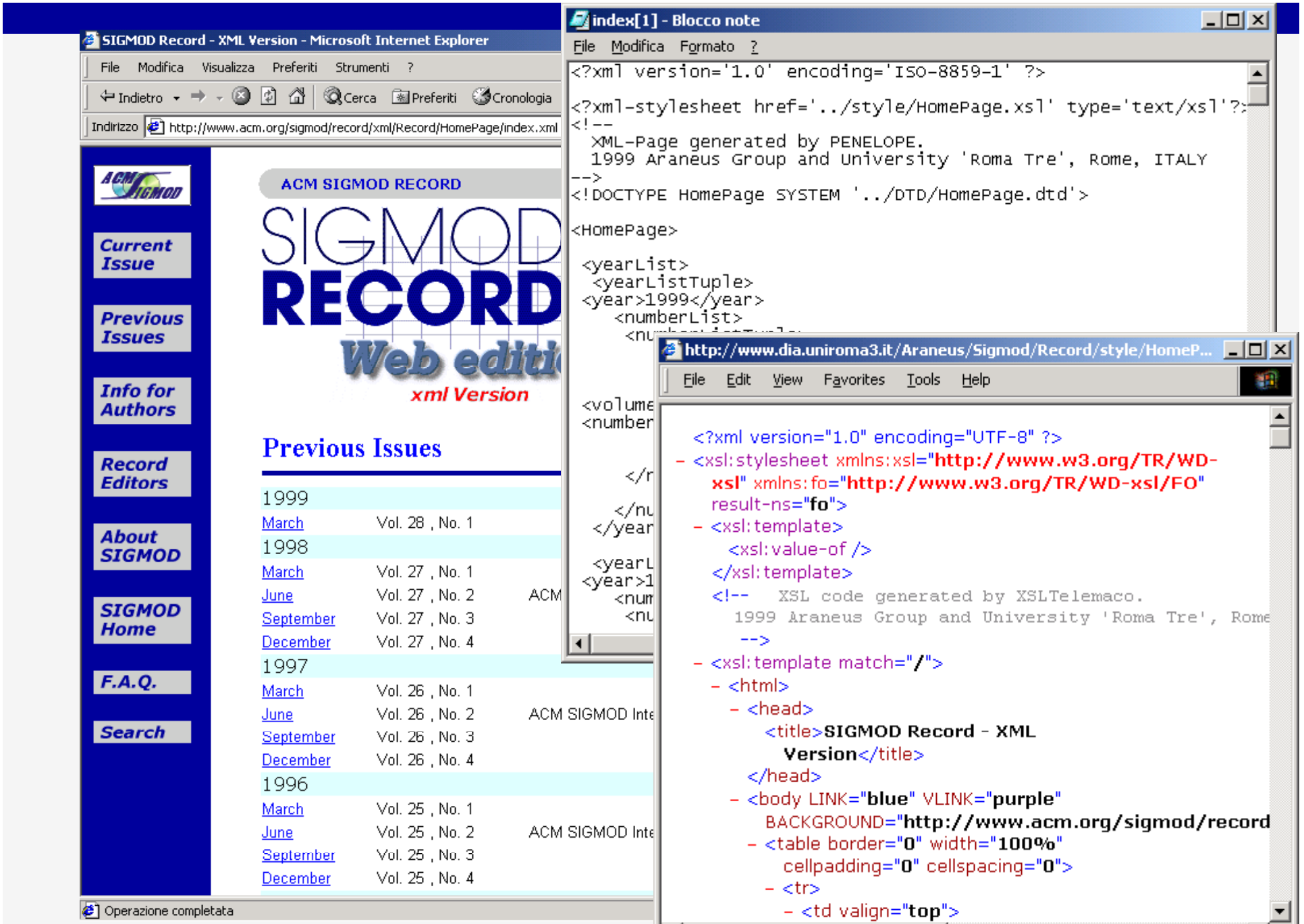
Integration of Web Sites: The Integrated Web Museum

- Data are re-organized:
  - Uffizi, paintings organized by rooms
  - Louvre, Capodimonte, works organized by collections
  - Integrated Museum, organized by author

Web Site Re-Engineering: ACM Sigmod Record in XML

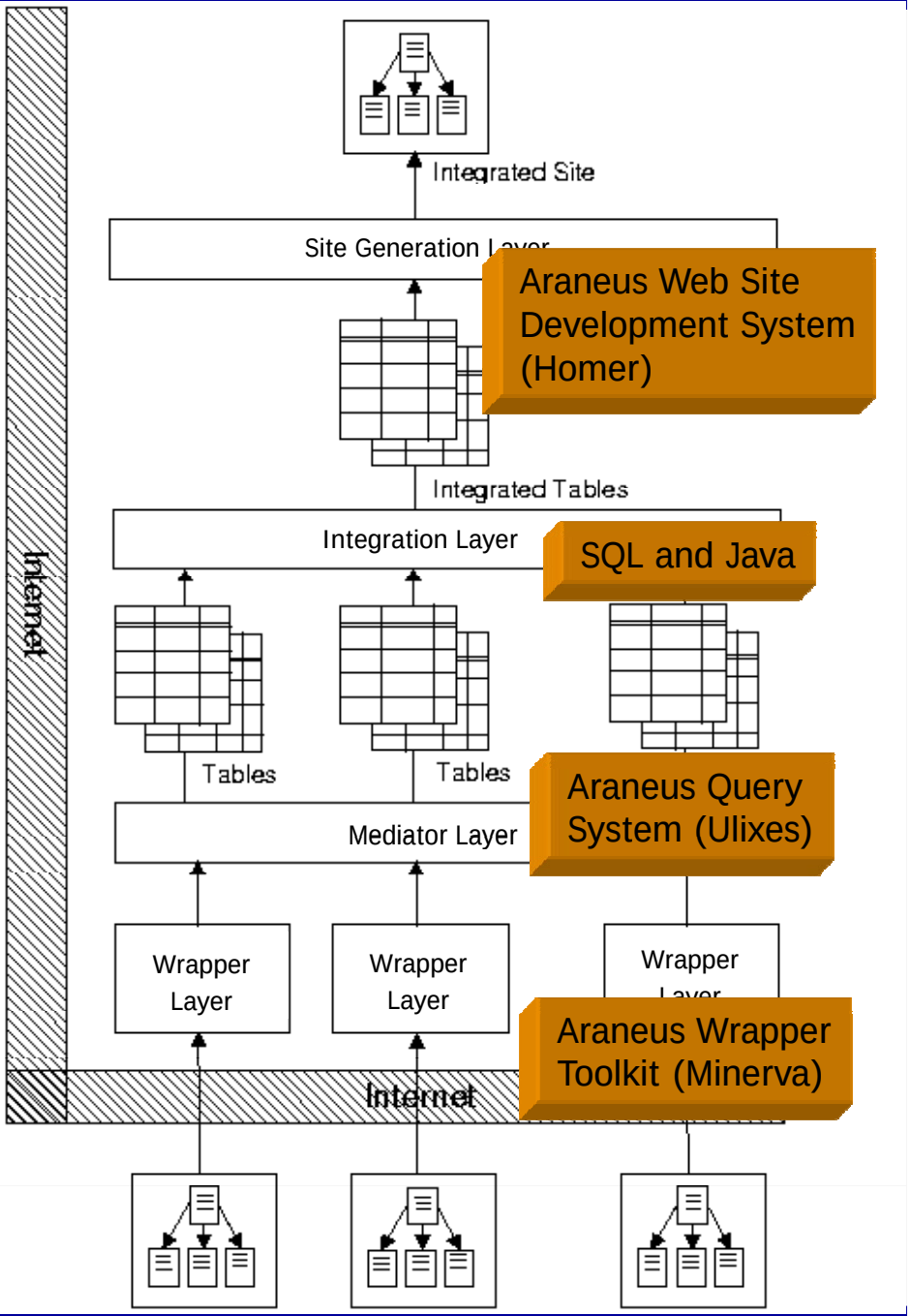
- The ACM Sigmod Record Site (acm.org/sigmod/record)
  - an existing site in HTML
- The ACM Sigmod Record Site, XML Edition
  - a new site in XML with XSL Stylesheets


SIGMOD Record - Web edition / Previous Issues - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Cerca Preferiti Cronologia
Indirizzo http://www.acm.org/sigmod/record/previous.html
Current Issue
Previous Issues
Info for Authors
Record Editors
About SIGMOD
SIGMOD Home
F.A.Q.
Search
Volume 29 Number 1
SIGMOD RECORD
Web edition
Previous Issues
1999
December 1999 Vol 28, No 4
September 1999 Vol 28, No 3
June 1999 Vol 28, No 2 (SIGMOD 1999 Proceedings)
March 1999 Vol 28, No 1 (Semantic Interoperability in Global Information Systems)
1998
December 1998 Vol 27, No 4 (Electronic Commerce)
September 1998 Vol 27, No 3
June 1998 Vol 27, No 2 (SIGMOD 1998 Proceedings)
March 1998 Vol 27, No 1
1997
December 1997 Vol 26, No 4 (Management of semi-structured data)
September 1997 Vol 26, No 3
June 1997 Vol 26, No 2 (SIGMOD 1997 Proceedings)
March 1997 Vol 26, No 1 (Environment Information Systems)
1996
Operazione completata
Internet
previous[1] - Blocco note
File Modifica Formato ?
<html>
<head>
<title>SIGMOD Record - Web edition / Previous Issues</titl
</head>
<body LINK="blue" VLINK="purple" BACKGROUND="/sigmod/recor
<a name="topofpage">
<!-- ****** HEADER ************************************
<table border=0 width=100% cellpadding=0 cellspacing=0>
<tr> <td width=84 valign=top align=left>
<!-- ********* NAVIGATION BAR ON LEFT SIDE ***********
<a href="http://www.acm.org/sigmod" target=main><IMG
SRC="/sigmod/record/images/sigmod-logo.gif" WIDTH=84
<!-- ********* GUTTER space in the middle ************
<td width=24><IMG SRC="/sigmod/record/images/pixel_tr
<td valign=top>
<table border=0 cellpadding=0 cellspacing=0 width=100
<tr> <td bgcolor=white valign=middle align=right><IMG
<td bgcolor=#CCCCCC nowrap valign=middle>
<font face=verdana,times color=#0000CC s
Volume 29</td>
<td bgcolor=#CCCCCC nowrap valign=middle>
<IMG SRC="/sigmod/record/images/pixel_tr


SIGMOD Record - XML Version - Microsoft Internet Explorer
File Modifica Visualizza Preferiti Strumenti ?
Indietro Cerca Preferiti Cronologia
Indirizzo http://www.acm.org/sigmod/record/xml/Record/HomePage/index.xml
ACM SIGMOD RECORD
SIGMOD RECORD
Web editi
xml Version
Current Issue
Previous Issues
Info for Authors
Record Editors
About SIGMOD
SIGMOD Home
F.A.Q.
Search
Previous Issues
1999
March Vol. 28 , No. 1
1998
March Vol. 27 , No. 1
June Vol. 27 , No. 2
September Vol. 27 , No. 3
December Vol. 27 , No. 4
1997
March Vol. 26 , No. 1
June Vol. 26 , No. 2
September Vol. 26 , No. 3
December Vol. 26 , No. 4
1996
March Vol. 25 , No. 1
June Vol. 25 , No. 2
September Vol. 25 , No. 3
December Vol. 25 , No. 4
Operazione completata
index[1] - Blocco note
File Modifica Formato ?
<?xml version='1.0' encoding='ISO-8859-1' ?>
<?xml-stylesheet href='../style/HomePage.xsl' type='text/xsl'?>
<!--
XML-Page generated by PENELOPE.
1999 Araneus Group and University 'Roma Tre', Rome, ITALY
-->
<!DOCTYPE HomePage SYSTEM '../DTD/HomePage.dtd'>
<HomePage>
<yearList>
<yearListTuple>
<year>1999</year>
<numberList>
http://www.dia.uniroma3.it/Araneus/Sigmod/Record/style/HomeP...
File Edit View Favorites Tools Help
<?xml version="1.0" encoding="UTF-8" ?>
- <xsl:stylesheet xmlns:xsl="http://www.w3.org/TR/WD-xsl" xmlns:fo="http://www.w3.org/TR/WD-xsl/FO" result-ns="fo">
- <xsl:template>
<xsl:value-of />
</xsl:template>
<!-- XSL code generated by XSLTelemaco.
1999 Araneus Group and University 'Roma Tre', Rome
-->
- <xsl:template match="/">
- <html>
- <head>
<title>SIGMOD Record - XML Version</title>
</head>
- <body LINK="blue" VLINK="purple" BACKGROUND="http://www.acm.org/sigmod/record
- <table border="0" width="100%" cellpadding="0" cellspacing="0">
- <tr>
- <td valign="top">

The Araneus Approach

- Identification of sites of interest
- Wrapping of sites to extract information
- Navigation of site ad extraction of data
- Integration of data
- Generation of new sites



Building Applications in Araneus

- Phase A:
  Reverse Engineering Existing Sites
- Phase B:
  Data Integration
- Phase C:
  Developing New Integrated Sites

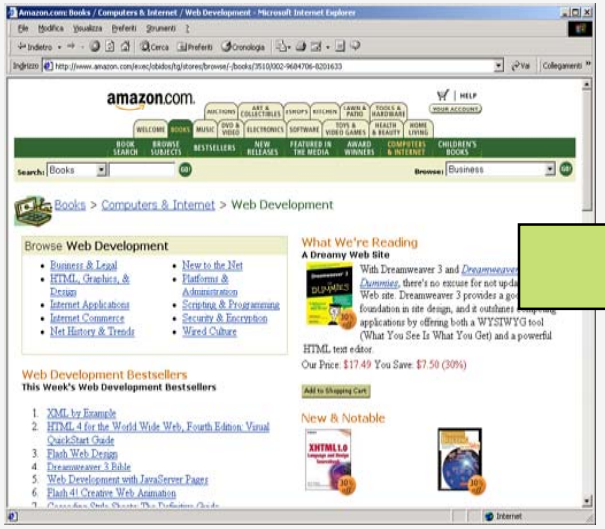
Building Applications in Araneus
Phase A: Reverse Engineering

- First Step: Deriving the logical structure of data in the site → ADM Scheme
- Second Step: Wrapping pages in order to map physical HTML sources to database objects
- Third Step: Extracting Data from the Site by Queries and Navigation

Information Extraction Task

- Information extraction task
  - source format: plain text with HTML markup (no semantics)
  - target format: database table or XML file (adding structure, i.e., “semantics”)
  - extraction step: parse the HTML and return data items in the target format
- “Wrapper”
  - piece of software designed to perform the extraction step

Notion of Wrapper



HTML page

Wrapper

BookTitle	Author	Editor
The HTML Sourcebook	**J. Graham**	**...**
Computer Networks	**A. Tannenbaum**	**...**
Database Systems	**R. Elmasri, S. Navathe**	**...**
Data on the Web	**S. Abiteboul, P.Buneman, D.Suciu**	**...**

database table(s)
(or XML docs)

Intuition:
use extraction rules based on HTML markup

Intuition Behind Extraction

teamName	*town*
Atalanta	Bergamo
Inter	Milano
Juventus	Torino
Milan	Milano
…	…

```
<html><body>
<h1>Italian Football Teams</h1>
<ul>
<li><b>Atalanta</b> - <i>Bergamo</i> <br>
<li><b>Inter</b> - <i>Milano</i>     <br>
<li><b>Juventus</b> - <i>Torino</i>  <br>
<li><b>Milan</b> - <i>Milano</i>     <br>
</ul>
</body></html>
```

Wrapper Procedure:

Scan document for
While there are more occurrences
    scan until
    extract teamName between and
    scan until
    extract town between and
    output [teamName, town]

The Araneus Wrapper Toolkit

	Procedural Languages	Grammars
Advantages	• Flexible in presence of irregularities and exceptions • Allows document restructurings	• Concise and declarative • Simple in the case of strong regularity
Drawbacks	• Tedious even with document strongly structured • Difficult to maintain	• Not flexible in presence of irregularities • Allow for limited restructurings

The Araneus Wrapper Toolkit: Minerva + Editor

- Minerva is a Wrapper generator
- It couples a declarative, grammar-based approach with the flexibility of procedural programming:
  - grammar-based
  - procedural exception handling

Development of wrappers

- Effective tools: drag and drop, example based, flexible
- Learning and automation (or semiautomation)
- Maintenance support:
  - Web sites change quite frequently
  - Minor changes may disrupt the extraction rules
  - maintaining a wrapper is very costly
  - this has motivated the study of (semi)- automatic wrapper generation techniques

Observation

- Can XML or the Semantic Web help much? Probably not
  - the Web is a giant “legacy system” and many sites will never move to the new technology (we said this ten years ago and it still seems correct …)
  - many organizations will not be willing to expose their data in XML

A related topic: Grammar Inference

- The problem
  - given a collection of sample strings, find the language they belong to
- The computational model: identification in the limit [Gold, 1967]
  - a learner is presented successively with a larger and larger corpus of examples
  - it simultaneously makes a sequence of guesses about the language to learn
  - if the sequence eventually converges, the inference is correct

The RoadRunner Approach

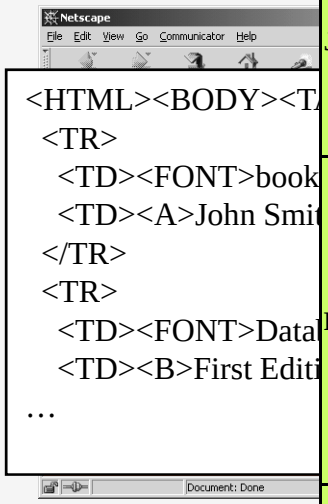
- RoadRunner
  - [Crescenzi, Mecca, Merialdo 2001] a research project on information extraction from Web sites
- The goal
  - developing fully automatic, unsupervised wrapper induction techniques
- Contributions
  - wrapper induction based on grammar inference techniques (suitably adapted)

The RoadRunner Approach

- The target
  - large Web sites with HTML pages generated by programs (scripts) based on the content of an underlying data store
  - the data store is usually a relational database
- The process
  - data are extracted from the relational tables and possibly joined and nested
  - the resulting dataset is exported in HTML format by attaching tags to values

The Schema Finding Problem

Wrapper Output



A	B				
	C	D	E		
			F	G	H
John Smith	Database Primer	This book …	First Edition, …	1998	20$
			Second Edition, …	2000	30$
	Computer Systems	An undergraduate …	First Edition, …	1995	40$
Paul Jones	XML at Work	A comprehensive …	First Edition, …	1999	30$
	HTML and Scripts	An useful HTML …	*null*	1993	30$
			Second Edition, …	1999	45$
	JavaScripts	A must in …	*null*	2000	50$
…	…	…	…	…	…

Wrapper

```
<HTML><BODY><TABLE>
 <TR>
  <TD><FONT>books.com<FONT></TD>
  <TD><A> #PCDATA</A></TD></TR>
 (<TR>
  ( <TD><FONT> #PCDATA <FONT></TD>
      ( <TD><B> #PCDATA </B> )? </TD>
  ( <TR><TD><FONT> #PCDATA <FONT>
      <B> #PCDATA </B> </TD></TR> )+
 )+…
```

Target Schema

```
SET (
 TUPLE (A : #PCDATA;
        B : SET (
            TUPLE ( C : #PCDATA;
                    D : #PCDATA;
                    E : SET (
                        TUPLE ( F: #PCDATA;
                                G: #PCDATA;
                                H: #PCDATA)))
```

The Schema Finding Problem

- Page class
  - collection of pages generated by the same script from a common dataset
  - these pages typically share a common structure
- Schema finding problem
  - *given a set of sample HTML pages belonging to the same class, automatically recover the source dataset*
  - *i.e., generate a **wrapper** capable of extracting the source dataset from the HTML code*

Contributions of RoadRunner

- An unsupervised learning algorithm for identifying prefix-markup languages in the limit
  - given a rich sample of HTML pages, the algorithm correctly identifies the language
  - the algorithm runs in polynomial time
  - from the language, when can infer the underlying schema
  - and then extract the source dataset

FLINT

Data Extraction and Integration from Imprecise Web Sources

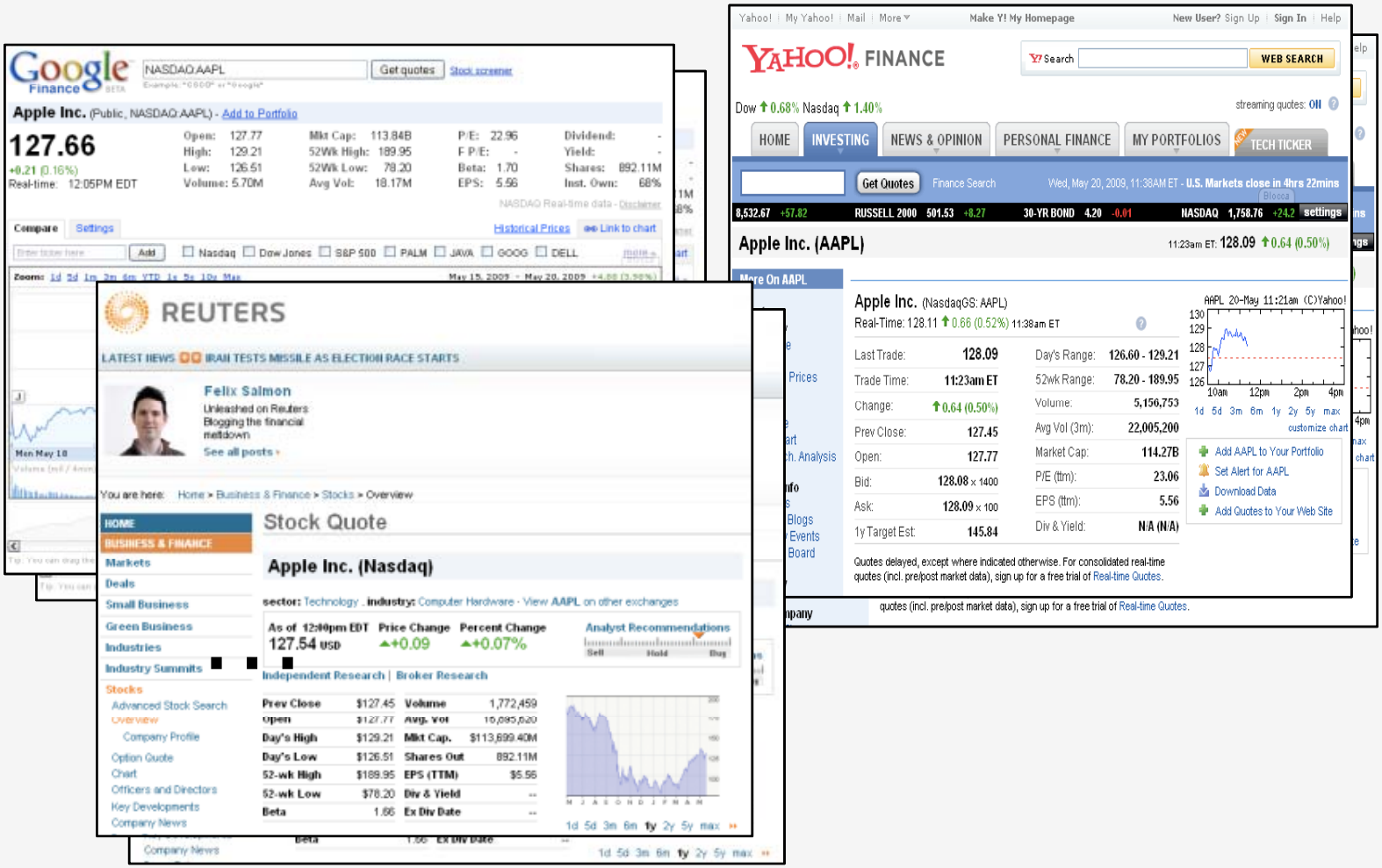
Lorenzo Blanco, Mirko Bronzi, Valter Crescenzi,
Paolo Merialdo, **Paolo Papotti**

Università degli Studi Roma Tre

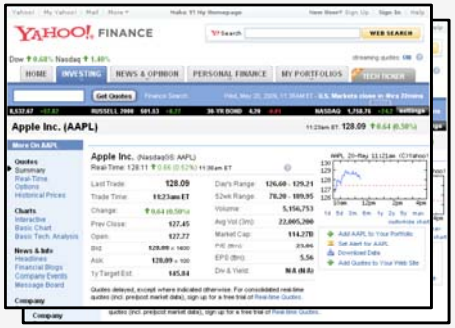
Flint

- A novel approach for the automatic extraction and integration of web data
- Flint aims at exploiting an unexplored publishing pattern that frequently occurs in data intensive web sites:
  - large amounts of data are usually offered by pages that encode one flat tuple for each page
  - for many disparate real world entities (e.g. stock quotes, people, travel packages, movies, books, etc) there are hundreds of web sites that deliver their data following this publishing strategy
  - These collections of pages can be though as HTML encodings of a relation (e.g. "stock quote" relation)

Example


Google Finance
NASDAQ:AAPL
Get quotes
Apple Inc. (Public, NASDAQ:AAPL) - Add to Portfolio
127.66
+0.21 (0.16%)
Real-time: 12:05PM EDT
Open: 127.77
High: 129.21
Low: 126.51
Volume: 5.70M
Mkt Cap: 113.84B
52Wk High: 189.95
52Wk Low: 78.20
Avg Vol: 18.17M
P/E: 22.96
F P/E: -
Beta: 1.70
EPS: 5.56
Dividend: -
Yield: -
Shares: 892.11M
Inst. Own: 68%
Historical Prices
Link to chart
REUTERS
LATEST NEWS IRAN TESTS MISSILE AS ELECTION RACE STARTS
Felix Salmon
Unleashed on Reuters
Blogging the financial meltdown
See all posts
You are here: Home › Business & Finance › Stocks › Overview
Stock Quote
Apple Inc. (Nasdaq)
sector: Technology . industry: Computer Hardware · View AAPL on other exchanges
As of 12:00pm EDT
127.54 USD
Price Change +0.09
Percent Change +0.07%
Analyst Recommendations
Independent Research | Broker Research
Prev Close $127.45
Open $127.77
Day's High $129.21
Day's Low $126.51
52-wk High $189.95
52-wk Low $78.20
Beta 1.66
Volume 1,772,459
Avg. Vol 10,085,820
Mkt Cap. $113,699.40M
Shares Out 892.11M
EPS (TTM) $5.56
Div & Yield --
Ex Div Date --
HOME
BUSINESS & FINANCE
Markets
Deals
Small Business
Green Business
Industries
Industry Summits
Stocks
Advanced Stock Search
Overview
Company Profile
Option Quote
Chart
Officers and Directors
Key Developments
Company News
YAHOO! FINANCE
WEB SEARCH
Dow 0.68% Nasdaq 1.40%
streaming quotes: Off
HOME
INVESTING
NEWS & OPINION
PERSONAL FINANCE
MY PORTFOLIOS
TECH TICKER
Get Quotes
Finance Search
Wed, May 20, 2009, 11:38AM ET - U.S. Markets close in 4hrs 22mins
8,532.67 +57.82
RUSSELL 2000 501.53 +8.27
30-YR BOND 4.20 -0.01
NASDAQ 1,758.76 +24.2
settings
Apple Inc. (AAPL)
11:23am ET: 128.09 0.64 (0.50%)
More On AAPL
Apple Inc. (NasdaqGS: AAPL)
Real-Time: 128.11 0.66 (0.52%) 11:38am ET
Last Trade: 128.09
Trade Time: 11:23am ET
Change: 0.64 (0.50%)
Prev Close: 127.45
Open: 127.77
Bid: 128.08 x 1400
Ask: 128.09 x 100
1y Target Est: 145.84
Day's Range: 126.60 - 129.21
52wk Range: 78.20 - 189.95
Volume: 5,156,753
Avg Vol (3m): 22,005,200
Market Cap: 114.27B
P/E (ttm): 23.06
EPS (ttm): 5.56
Div & Yield: N/A (N/A)
Add AAPL to Your Portfolio
Set Alert for AAPL
Download Data
Add Quotes to Your Web Site
Quotes delayed, except where indicated otherwise. For consolidated real-time quotes (incl. pre/post market data), sign up for a free trial of Real-time Quotes.

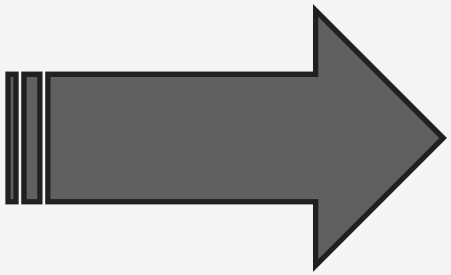
Flint goals

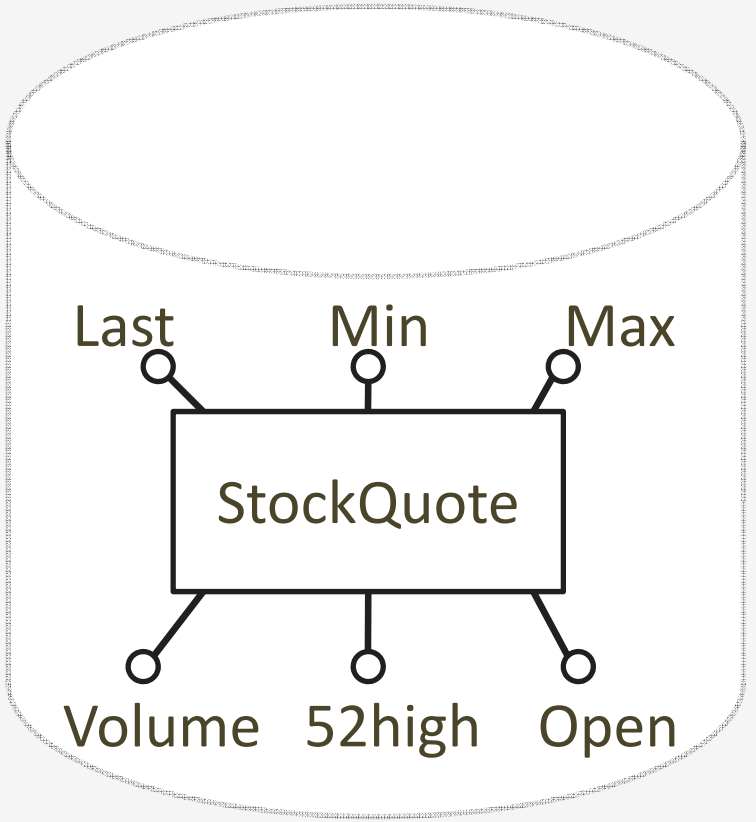

YAHOO! FINANCE
Apple Inc. (AAPL)


Google Finance
Apple Inc.
127.66

. . .

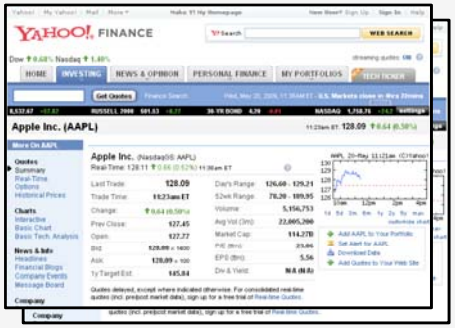

REUTERS
Stock Quote
Apple Inc. (Nasdaq)




Last
Min
Max
StockQuote
Volume
52high
Open

Flint goals

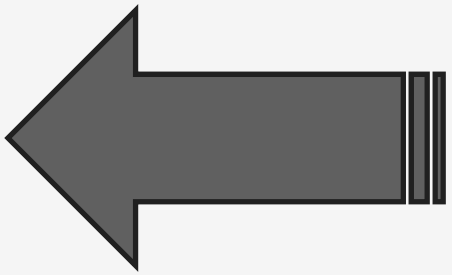
q_1



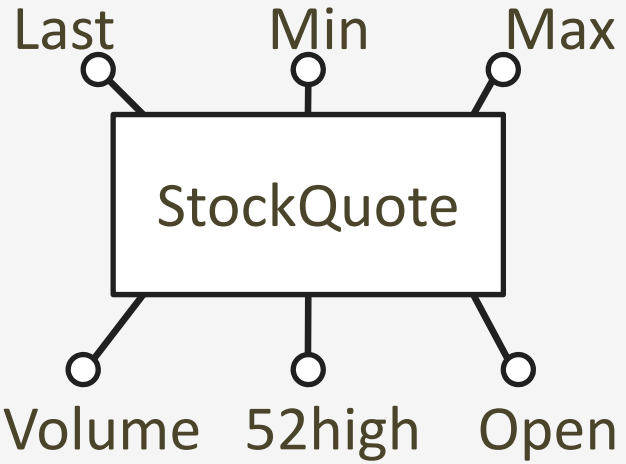
q_2

. . .

q_n



Q



Extraction and Integration

- Flint builds on the observation that although structured information is spread across a myriad of sources, the web scale implies a relevant redundancy, which is apparent
  - at the intensional level: many sources share a core set of attributes
  - at the extensional level: many instances occur in a number of sources
- The extraction and integration algorithm takes advantage of the coupling of the wrapper inference and the data integration tasks

Flint main components

- a crawling algorithm that gathers collections of pages containing data of interest
- a wrapper inference and data integration algorithm that automatically extracts and integrates data from pages collected by the crawler
- a framework to evaluate (in probabilistic terms) the quality of the extracted data and the reliability/trustiness of the sources

Crawling

- Given as input a small set of sample pages from distinct web sites
- The system automatically discovers pages containing data about other instances of the conceptual entity exemplified by the input samples

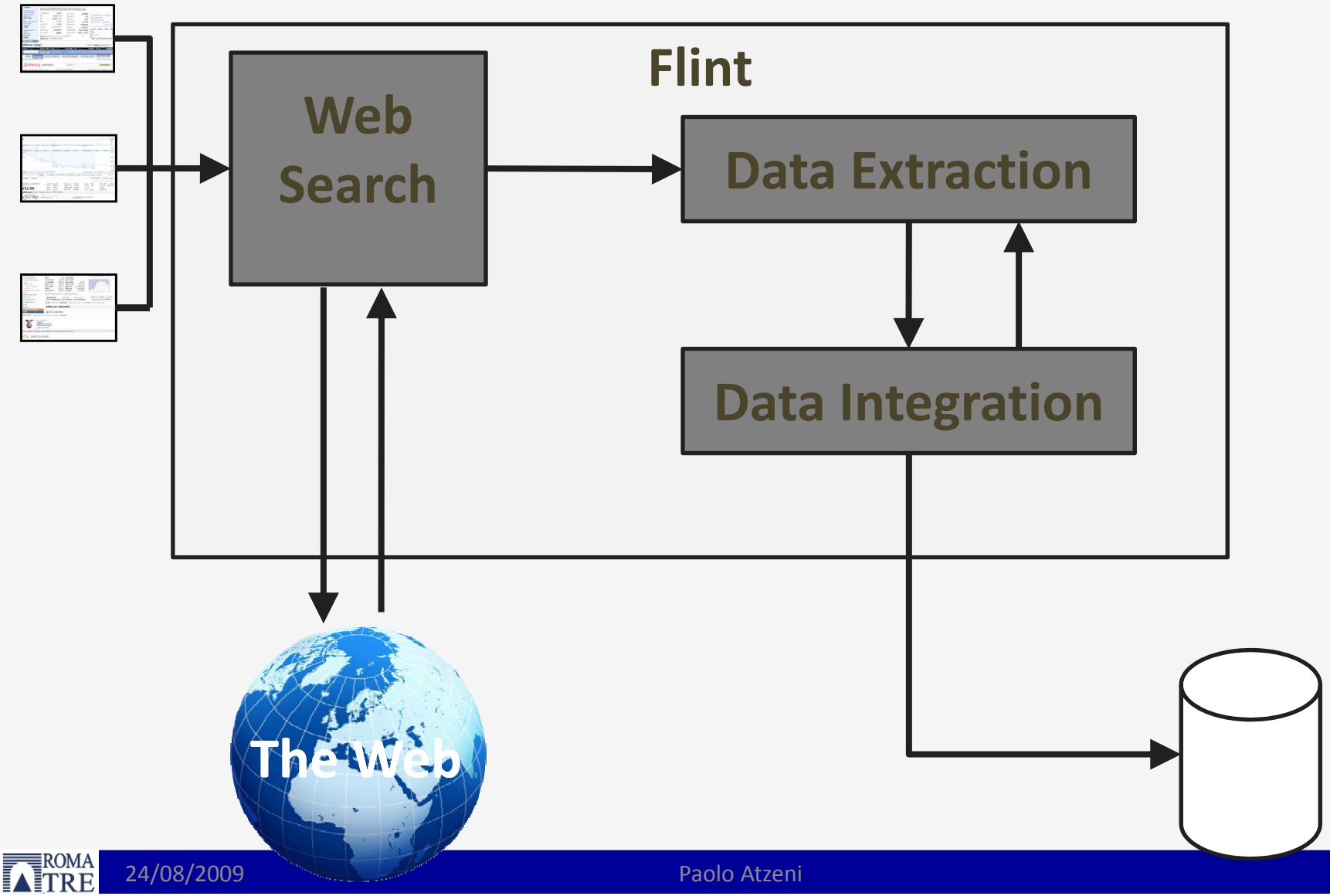
The crawler, approach

- Given a bunch of sample pages
- crawl the web sites of the sample pages to gather other pages offering the same type of information
- extract a set of keywords that describe the underlying entity
- do
  - launch web searches to find other sources with pages that contain instances of the target entity
  - analyze the results to filter out irrelevant pages
  - crawl the new sources to gather new pages
- while new pages are found

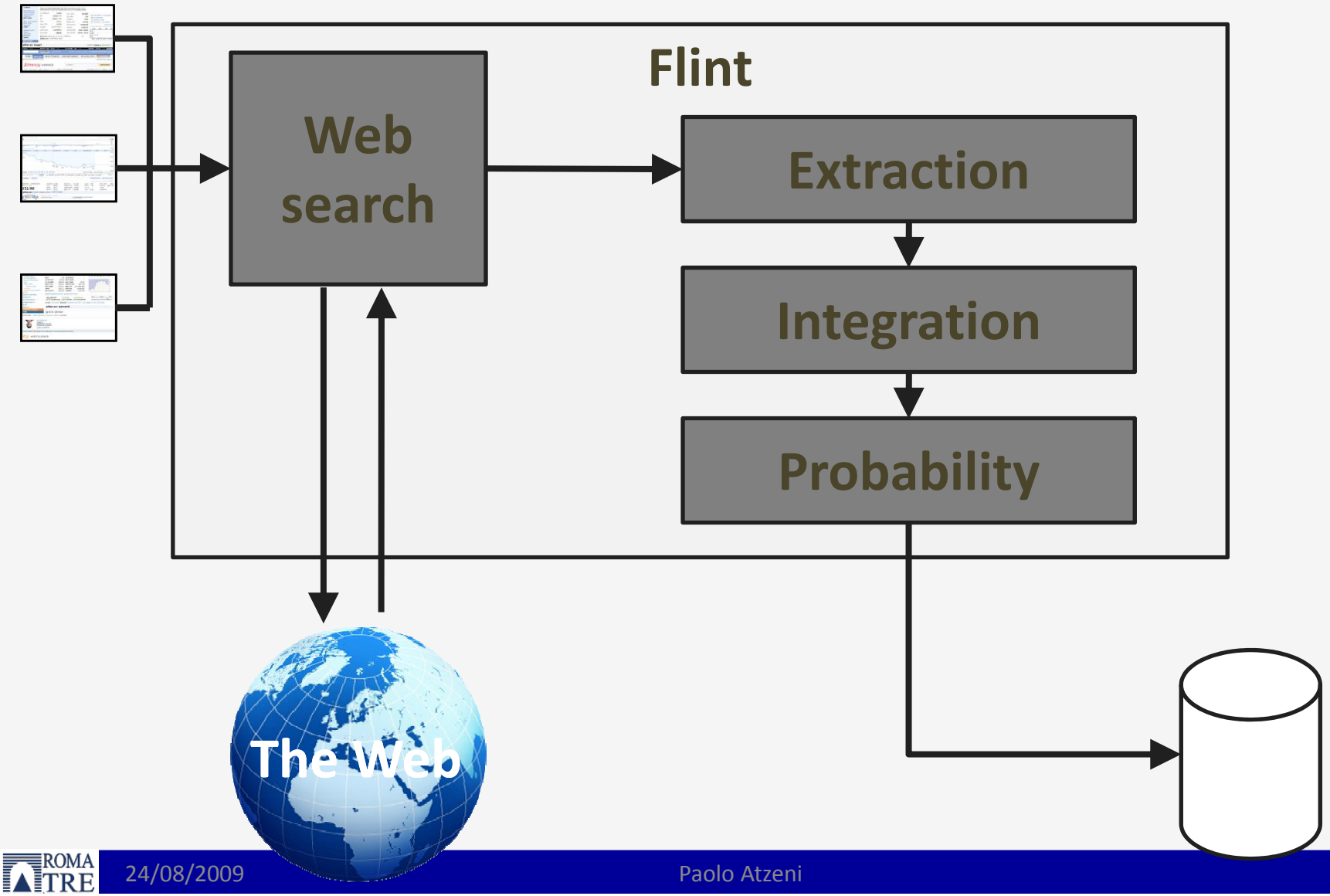
Data and Sources Reliability

- Redundancy among autonomous and heterogeneous sources implies inconsistencies and conflicts in the integrated data because sources can provide different values for the same property of a given object
- A concrete example: on April 21th 2009, the open trade for the Sun Microsystem Inc. stock quote published by the CNN Money, Google Finance, and Yahoo! Finance web sites, was 9.17, 9.15 and 9.15, respectively.

System architecture



System architecture



Summary

- The exploitation of structure is essential for extracting information from the Web
- In the Web world, task automation is needed to get scalability
- Interaction of different activities (extraction, search, transformation, integration) is often important