

YOCSEF长沙论坛:

Web Data Integration



孟小峰

中国人民大学信息学院

xfmeng@ruc.edu.cn

长沙, 10/2003



报告内容

- 引言
- 信息集成技术的发展
- Web数据集成方法
- Wrapper生成技术
- Wrapper维护问题
- 基于 Web Service 的 Web 数据集成系统
(WDIWS)
- 我们的工作及展望



The Web of yesterday

- Protocol: HTTP
- Documents: HTML
- Millions of independent Web sites and billions of documents
- Browsing and full-text indexing
- Publication of databases using forms
- Data management with the Web
 - HTML is primarily to be read by humans
 - Data management applications over Web data
 - Based on hand-made wrappers
 - Expensive, incomplete, short-lived, not adapted to the Web constant change

No real support for distributed data management!



Data on the Web

- Different formats: relational, metadata, documents, text
 - A Web standard for data exchange, XML, is fixing it
 - XML captures all kinds of information over a wide spectrum
 - XML comes with a family of emerging standards: XML schema, XSL/T, Xquery, domain specific schemas...
- Different computers, platforms, languages, applications
 - A standard for Web services, SOAP, is fixing it
 - SOAP allows ubiquitous computing on the Internet
 - SOAP comes with a family of emerging standards: WSDL, UDDI
- This provides a uniform access to information...
 - ...the dream for distributed data management



报告内容

- 引言
- 信息集成技术的发展
- Web数据集成方法
- Wrapper生成技术
- Wrapper维护问题
- 基于 Web Service 的 Web 数据集成系统
(WDIWS)
- 我们的工作及展望



信息集成技术

- 信息系统集成技术已经经历了二十多年的发展过程，研究者已提出了很多信息集成的体系结构和实现方案，然而这些方法研究的主要集成对象是传统的异构数据库系统。随着Internet的飞速发展，网络迅速成为一种重要的信息传播和交换的手段，尤其是Web上，有着及极其丰富的数据来源。如何获取Web上的有用数据并加以综合利用，即构建Web信息集成系统，成为一个引起广泛关注的研究领域。



信息集成技术发展

- 单个的联邦系统：

- 将所有数据源统一到一个单一的集成系统中。这种方法比较简单，集成系统有统一的数据模式，不用考虑分布数据的转化和统一。但是，它存在一系列的问题：首先，构建这样一个集中式的系统需要很长的开发时间，要求高性能的主机设备，实现代价较高；其次，系统的扩展和维护会涉及到整个系统，而且一个集成系统无法共享另一个集成系统的模块。



信息集成技术发展

■ 基于组件的分布式集成系统

- 用分布式的对象模型，诸如，微软的分布式组件对象模型(DCOM)、CORBA或Sun的RMI来构建信息集成系统。这种方法有效的避免了单个联邦系统带来的开发代价大，代码难以重用的问题，利用网络计算环境可以有效的实现复杂的大规模的信息集成。但是，DCOM，CORBA或RMI要求服务客户端与系统提供的服务本身之间必须进行紧密耦合，即要求一个同类基本结构。这样的系统往往十分脆弱：如果一端的执行机制发生变化，那么另一端便会崩溃。例如，如果服务器应用程序的接口发生更改，那么客户端便会崩溃。



信息集成技术发展

■ 基于Web Services的信息集成系统

- Internet的迅速普及和广泛应用对计算机技术的发展产生了深刻影响，桌面应用正在向网络应用转移，从网上获得的不仅是信息，还包括程序、交互式应用（即服务），操作界面将在浏览器层面上得到统一，兼容性由网络标准技术实现（如SOAP，UDDI，WSDL等）。在Web Services的框架下，使用一组Web Services协议，构建信息集成系统。对每个数据源都为其创建一个Web Service，然后使用WSDL向服务中心注册。当要构建一个新的集成应用时，集成端首先向注册中心发送查找请求收集并选择合适的数据源，然后通过SOAP协议从这些数据源获取数据。这种方法克服了上述两种方法的缺陷，具有完好封装，松散耦合，规范协议高度可集成能力等特性。因此，基于Web Services的信息集成方案是构建Web数据集成系统较为理想的体系结构。



报告内容

- 引言
- 信息集成技术的发展
- Web数据集成方法
- Wrapper生成技术
- Wrapper维护问题
- 基于 Web Service 的 Web 数据集成系统
(WDIWS)
- 我们的工作及展望



Web Data Integration

- 为用户自动或半自动地从web上获取数据建立多数据源的统一视图,并为用户提供有效的查询和信息发布方式
- Applications: catalogs (seeing products from many suppliers), digital libraries, scientific databases, enterprise-wide information resources, e-government, etc., etc.



Solutions

- Web query languages
 - WebSQL, W3QL, WebOQL, StruQL ...
- Wait for XML to emerge
 - Interoperation/Standards?
 - XML query language?
- Wrappers
 - Hand-written or semi-automatically generated parsers
 - Specific to source site, subject to change



Two Approaches

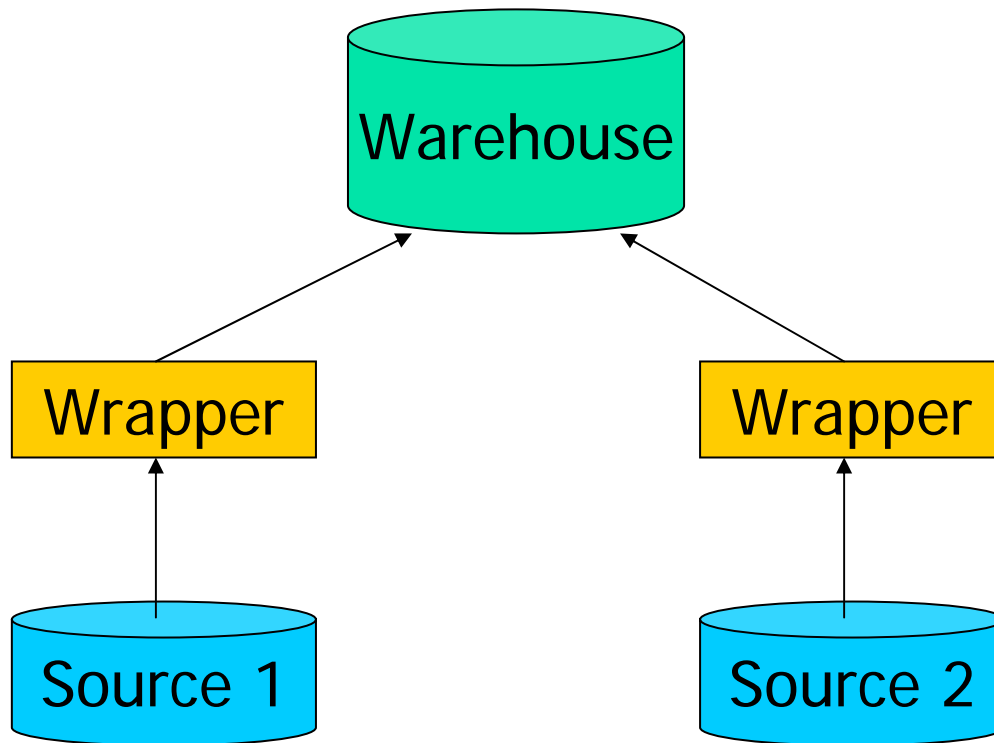
1. *Warehousing* :

- Collect data from sources into a “warehouse” periodically.
- Do queries at the warehouse, while the sources execute transactions invisibly.

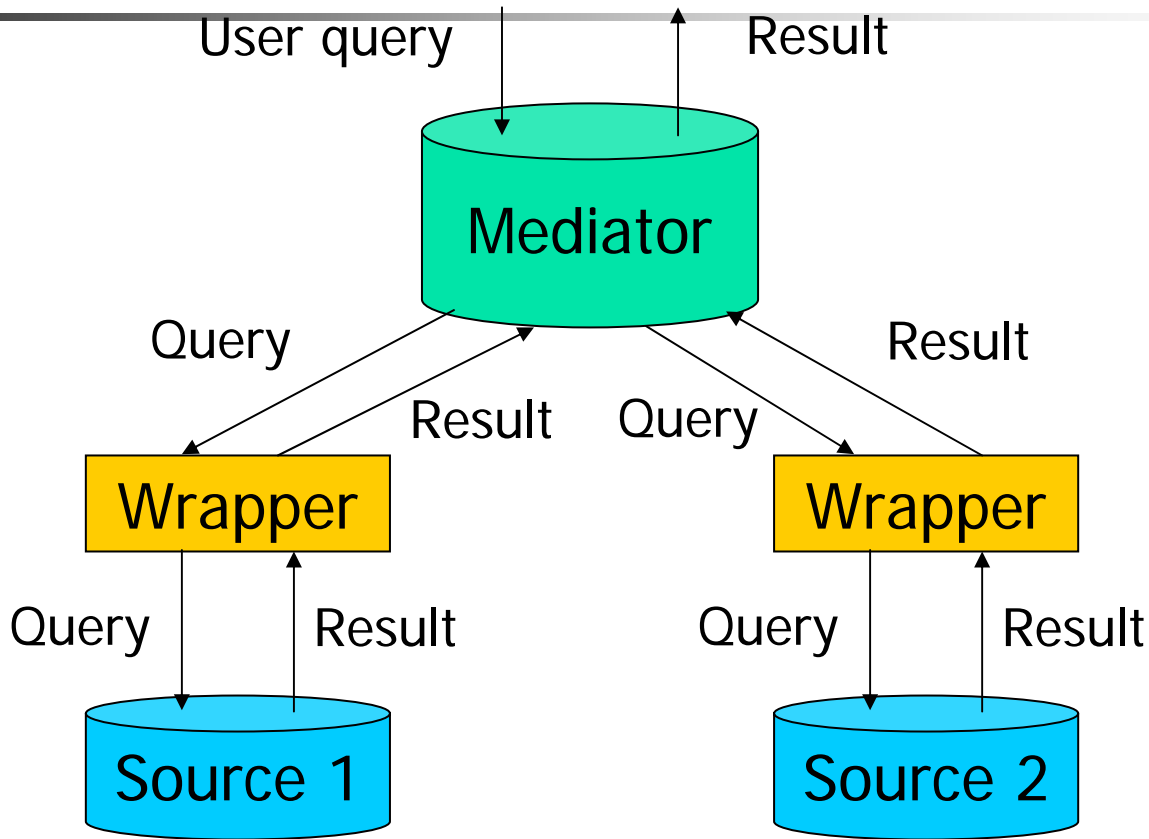
2. *Mediation* :

- Virtual warehouse processes queries by translating between common schema and local schemas at sources.

Warehouse Diagram



A Mediator





Two Mediation Approaches

1. *Query-centric* : Mediator processes queries into steps executed at sources.
2. *View-centric* : Sources are defined in terms of global relations; mediator finds all ways to build query from views.



Comparison

- Query-centric is simpler to implement.
 - Lets you have control of what the mediator does.
- View centric is more extensible.
 - Same query engine works for any number of sources.
 - Add a new source simply by defining what it contributes as a view of the global schema.



Research Issues

- Optimization, optimization, optimization.
 - In query centric systems: how do we choose a plan?
 - E.g., is it better to ask about buses first, or disks?
 - In view-centric systems, how do we select a sufficient set of solutions to get most or all of the possible answers?



报告内容

- 引言
- 信息集成技术的发展
- Web数据集成方法
- Wrapper生成技术
- Wrapper维护问题
- 基于 Web Service 的 Web 数据集成系统
(WDIWS)
- 我们的工作及展望



Wrapper

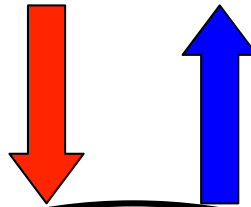
- Typical Wrapper
 - a procedure that is designed for extracting content of a particular information source and delivering the content of interesting in a self-describing representation (eg. XML).

Wrappers

- Wrapper for Web page

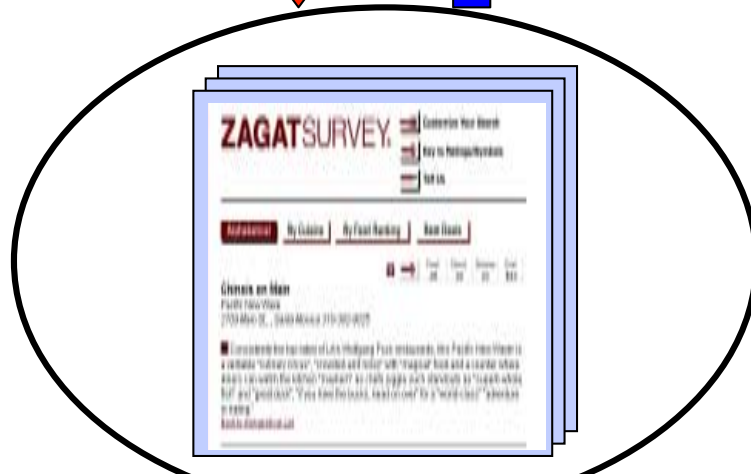
- provide uniform mechanism for extracting data from semi-structured sources (HTML, text, ...)
- transform semi-structured sources into structured

**Restaurants in
Santa Monica?**



Name	Address
Chinois on Main	2709 Main St.
Chao Dara	13 Union Sq.
?	...

Wrapper

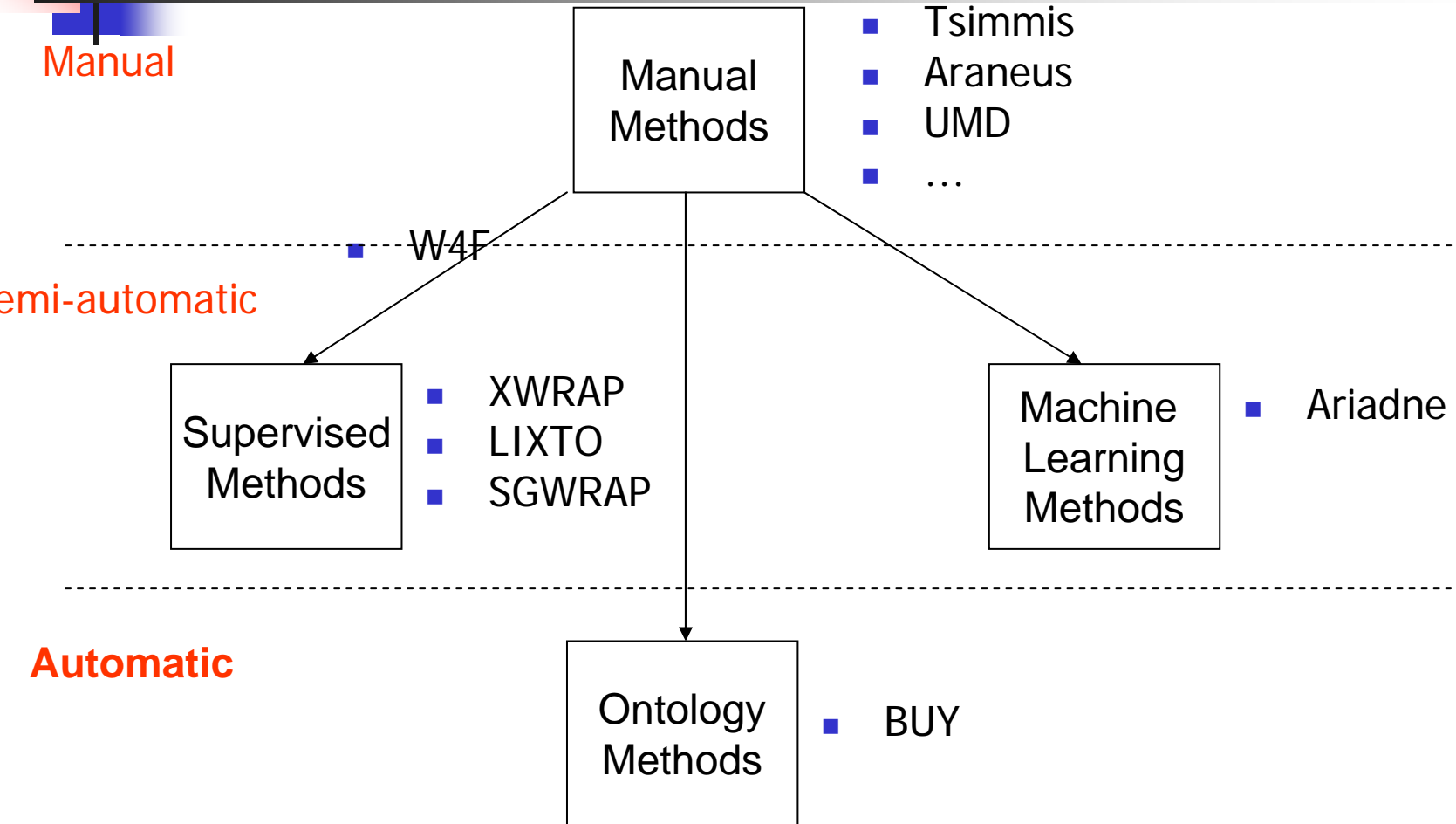




Why build wrappers?

- The ability enhanced to obtain the relevant information from an individual source
- An integrated access to several sources and a database-like query using a common language.

Wrappers - History





Wrapper生成技术

- Wrapper的生成方法可以分为三类
 - **Wrapper程序语言方法**
 - **机器学习的方法**
 - **受指导的交互式wrapper生成方法**
 - XWrap[18]使用程序化的规则体系并提供了有限的模式定义表达能力。
 - Lixto[12]提供了可视化的方式进行wrapper生成，用户可以通过浏览的方式来标记文档。
 - 文献[19,20]提出了模式导航的wrapper生成方法（SG-WRAP）



SGWRAP - Overview

- Schema Guided Wrapper Generator
- Features
 - Generating extraction rules with the guidance of user-defined schema
 - The wrapper generated based on the rules could be more accurate and better reflect the users requirements.
 - Using different schemas, the wrapper can be easily integrated into the different data integration process.

SGWRAP - Architecture

Schema Information

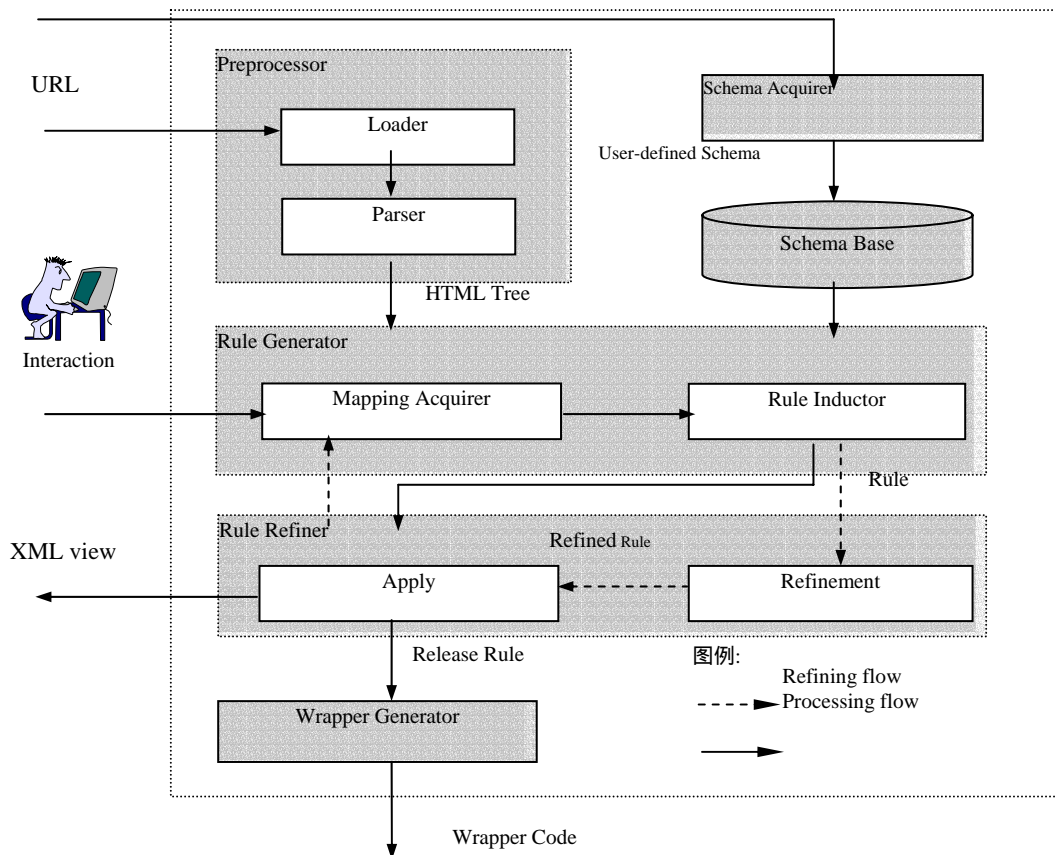


图4 SG-WRAP: 系统体系结构



报告内容

- 引言
- 信息集成技术的发展
- Web数据集成方法
- Wrapper生成技术
- Wrapper维护问题
- 基于 Web Service 的 Web 数据集成系统 (WDIWS)
- 我们的工作及展望



Wrapper维护问题

- Web设计者经常调整页面的格式
- 用户对页面的变化没有控制权
- 页面变化使wrapper失效
- 页面经常变化需要重新建立wrapper



维护的相关工作

- ❖ Kushmerick: 通过回归测试进行变化检测
- ❖ Knoblock: 获取抽取规则中的内容特征
- ❖ Cohen: 使用信息检索中文本相似方法重新定位数据项
- ❖ Chidlovskii: 将语法特征和内容特征作为分类的标准，对多页面进行多种分类和多遍扫描
- ❖ SG-WRAM: 提出基于模式的wrapper维护方法

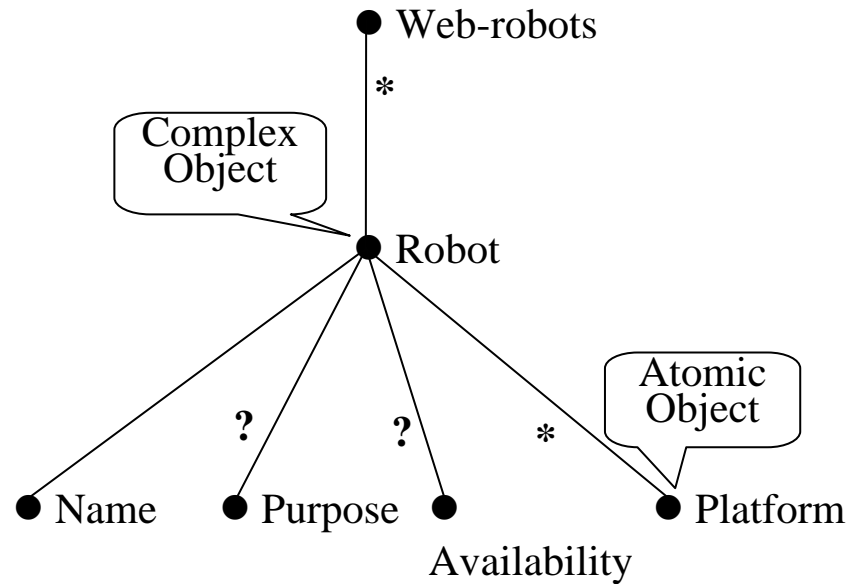


Preliminaries

- Schema tree
- HTML Tree
- Minimum semantic block

Schema tree

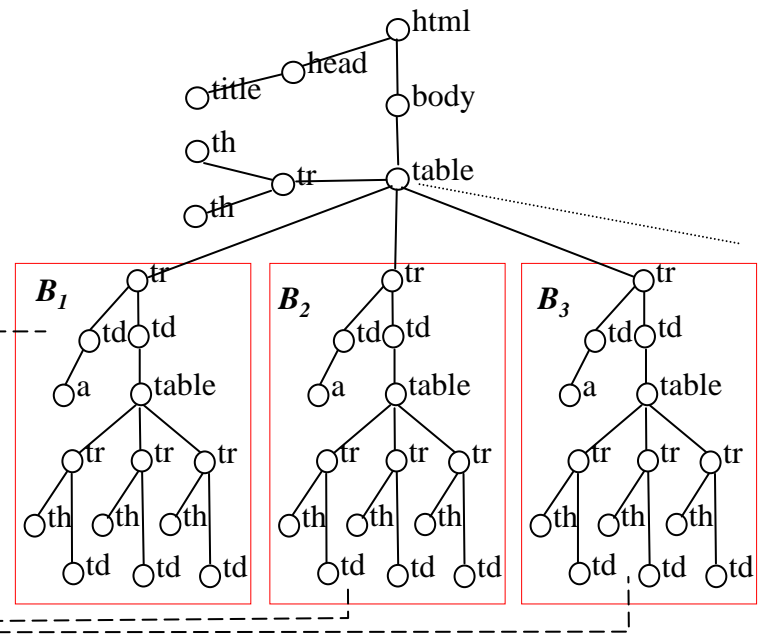
Name	Details
Acme Spider	Purpose: indexing maintenance statistics Availability: source Platform: java
Ahoz! The Homepage Finder	Purpose: maintenance Availability: none Platform: UNIX
Alkaline	Purpose: indexing Availability: binary Platform: unix windows95 windowsNT



HTML Tree & Semantic block

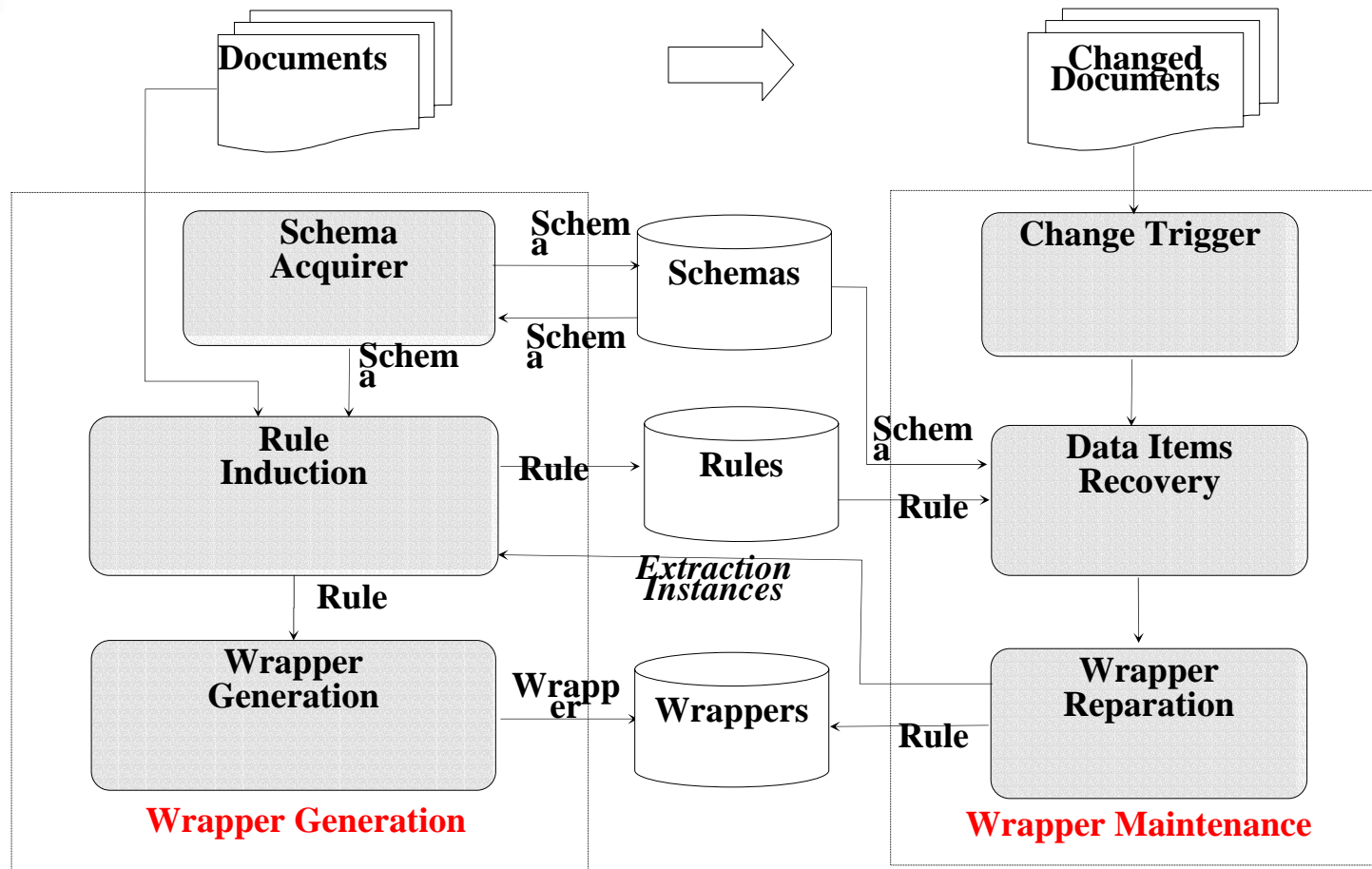
Name	Details
Acme Spider	Purpose: indexing maintenance statistics Availability: source Platform: java
Ahoy! The Homepage Fender	Purpose: maintenance Availability: none Platform: UNIX
Alkalne	Purpose: indexing Availability: binary Platform: unix windows95 windowsNT

(a) Source Pages of Web Robots



(b) HTML Tree

SGWRAM - Architecture





SGWRAM - Architecture

- 我们设计的wrapper维护由四个步骤组成：
 - *Data-feature discovery*: 数据特征从给定的DTD，以前的抽取规则和先前的抽取结果中计算得出；
 - *Data-item recovery*: 数据特征用于在新页面中识别有关的数据项；
 - *Block configuration*: 根据用户提供的模式信息，在HTML树结构上划分为若干子树，每个子树由符合模式定义的相关数据项组成，称为一个语义块；
 - *Wrapper reparation*: 从语义块中选择代表性的实例，对改变过的页面进行重新推理得出新的抽取规则。



报告内容

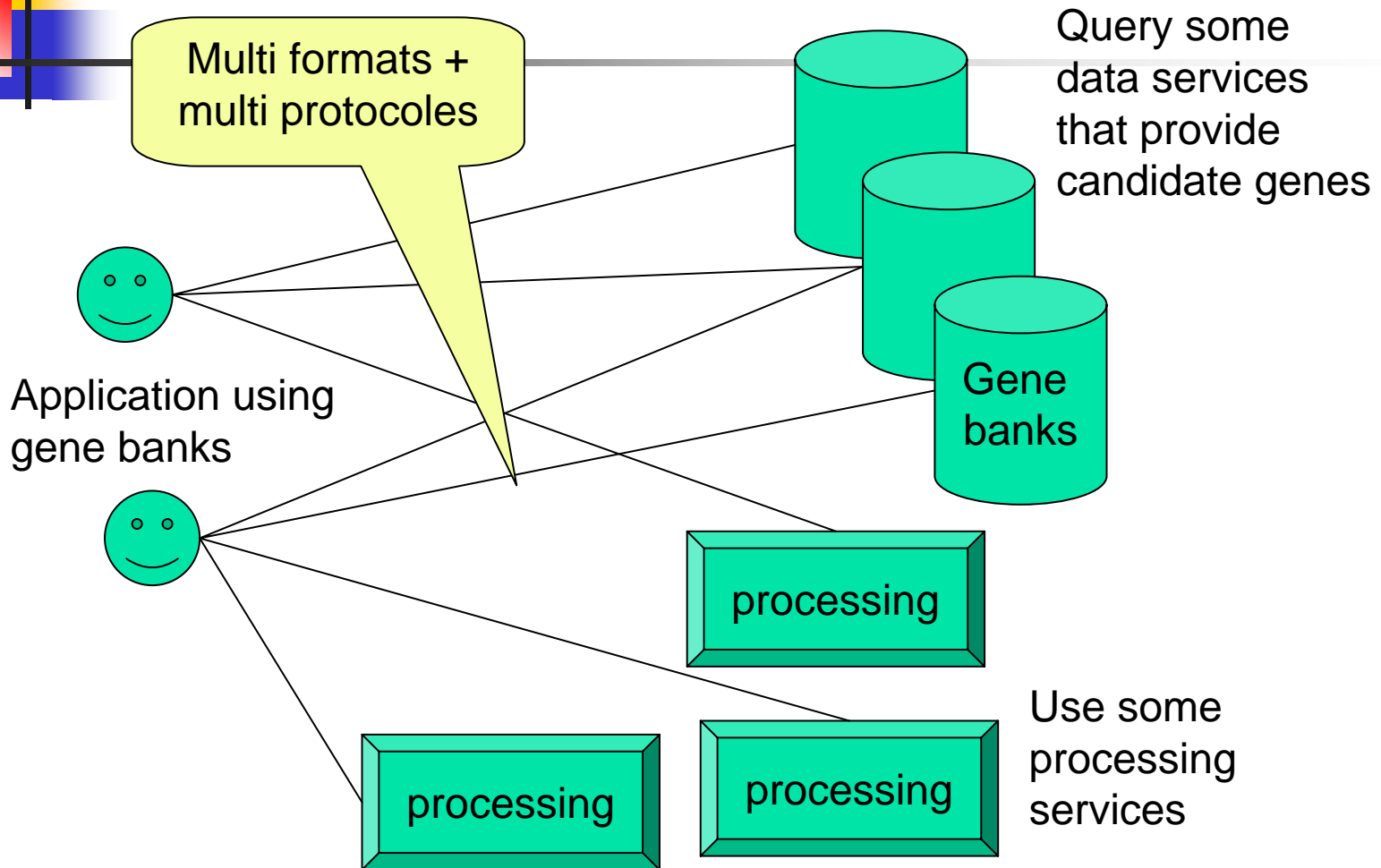
- 引言
- 信息集成技术的发展
- Web数据集成方法
- Wrapper生成技术
- Wrapper维护问题
- 基于 Web Service 的 Web 数据集成系统
(WDIWS)
- 我们的工作及展望



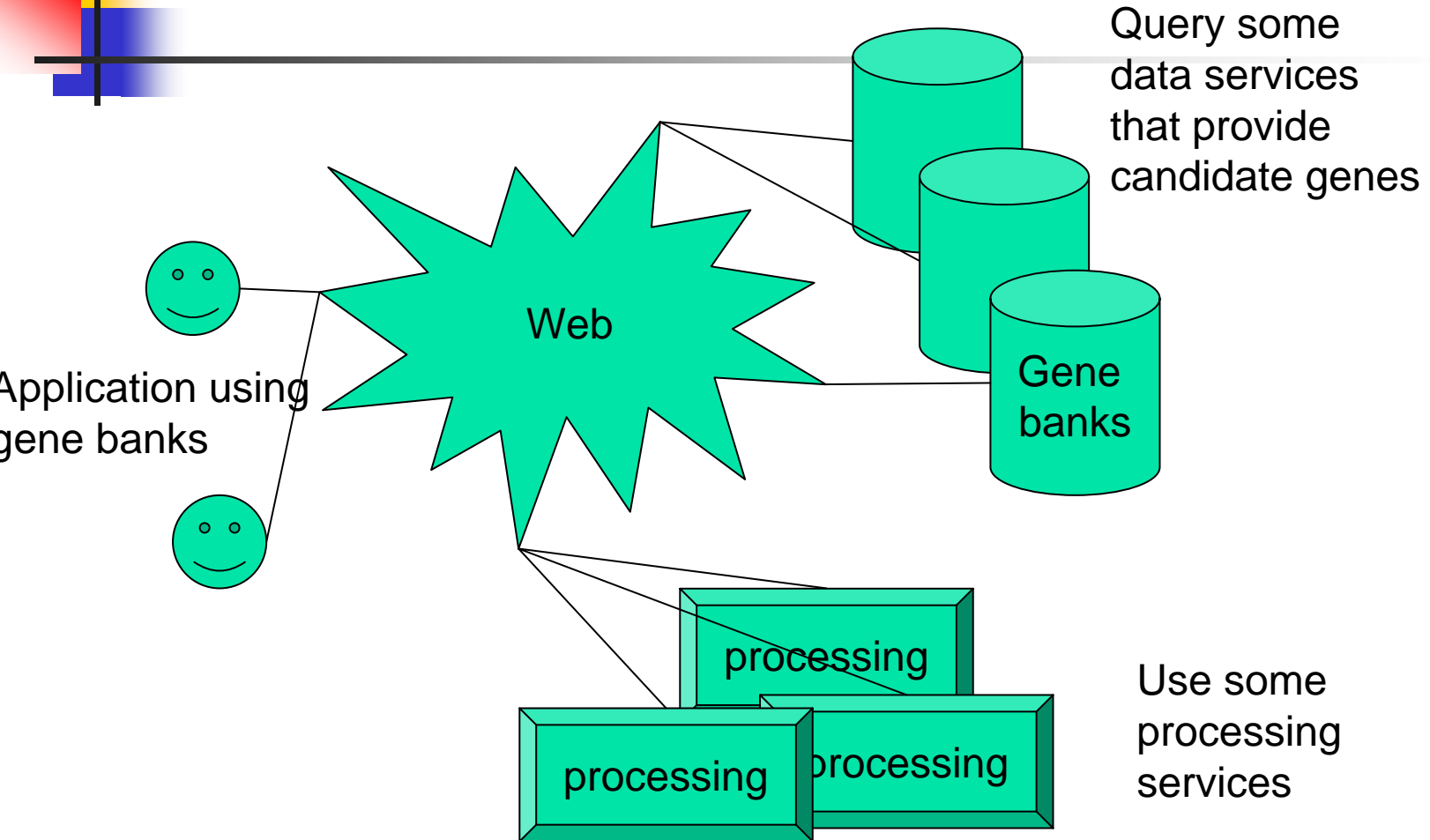
A standard for distributed computing: Web services

- Possibility to activate a method on some remote Web server
- Exchange information in XML: input and result are in XML
- ***Ubiquitous XML distributed computing infrastructure***
- 2 main application
 - E-Government
 - ***Access to remote data***
- With XML and Web services, it is possible
 - To get information from virtually anywhere
 - To provide information to virtually anywhere

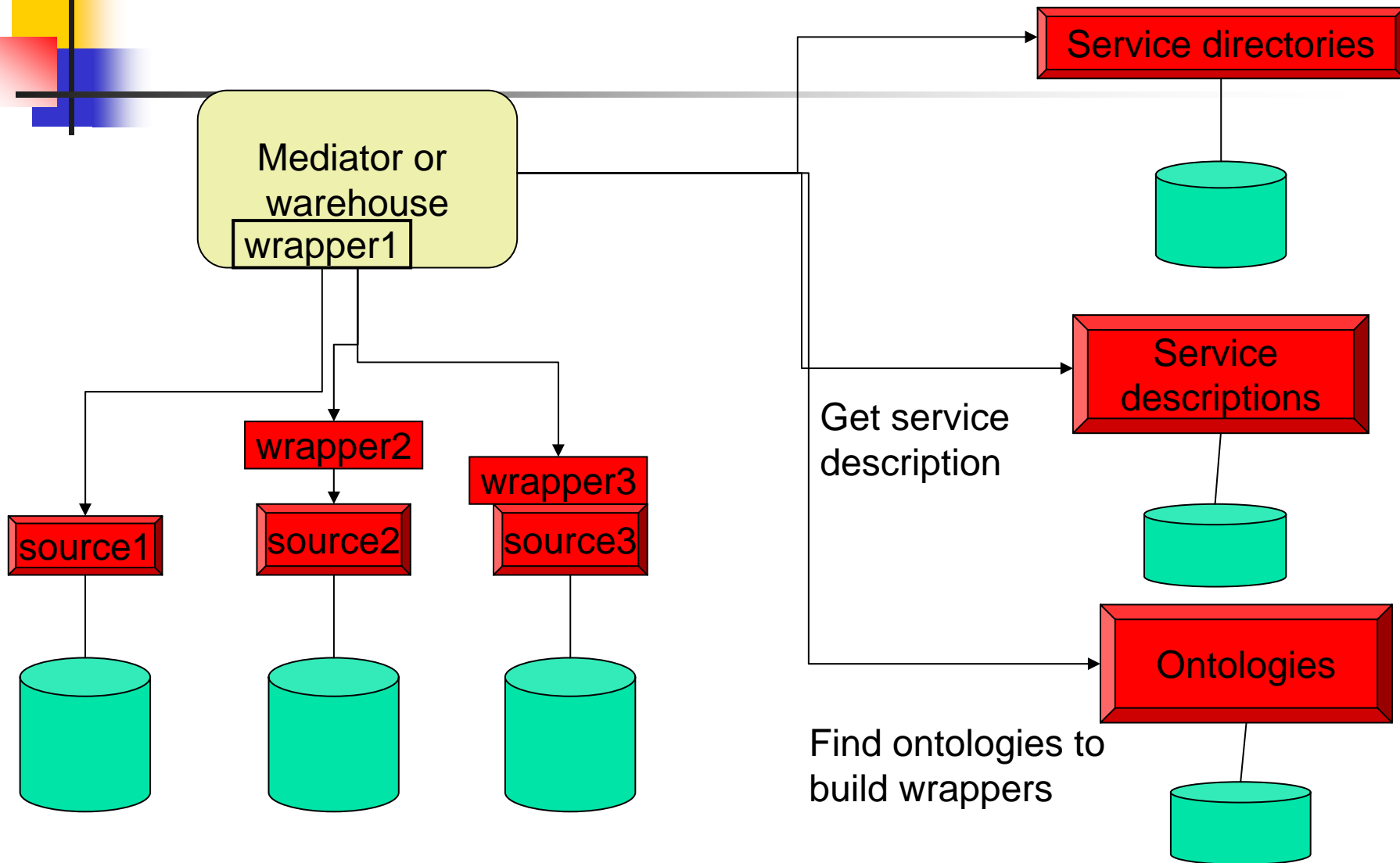
Accessing remote information



Same with Web services



Data integration – Logical view



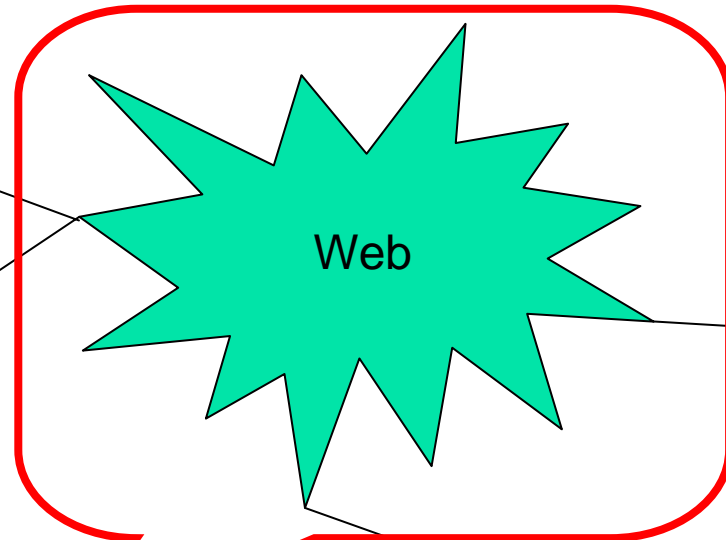
The Web service solution

Data and service repository

UDDI

RDF

Data and service semantics



Data and service description

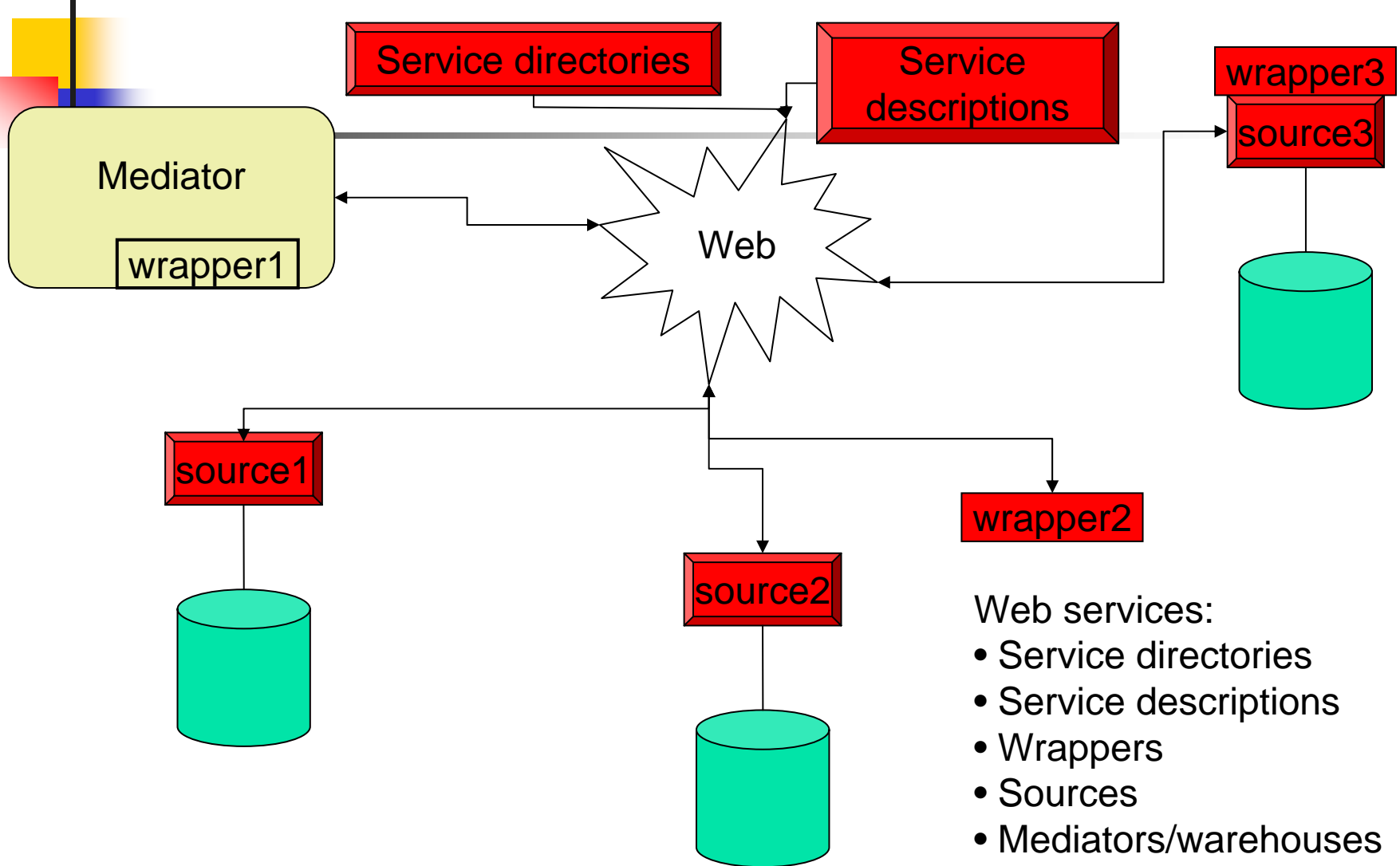
wsdl

worklow

wsfl

XML+SOAP

Mediation with Web services





Advantages for data integration

- A universal model for data integration = **XML**
 - Solves the heterogeneity issue
- A universal protocol for distribution = **SOAP**
- A language for describing the interface of data sources = **WSDL**
 - Simple object access protocol (something like Corba)
 - Web service description language (something like IDL)
 - Solves the interoperability issue
- A standard for publication and discovery of information = **UDDI**
 - Universal Description, Discovery and Integration
- A standard for describing the semantics of sources = **RDF**
 - Resource description framework



Advantages – continued – the goal

- The system can find a new source of information using UDDI
- Understand its syntax using WSDL
- Understand its semantics using RDF
- Get it using SOAP
- The information is in XML, can be restructured and integrated automatically
- Not yet... But soon?



基于Web Service的Web数据集成系统（WDIWS）

- 基于Web Services构建Web数据集成系统是目前较为理想的方法。
- WDIWS系统的初步目标是实现一个架构于Web Service技术之上的Web数据集成系统。
- 在这个系统中，用户需要提供数据源信息以用于确定数据源和查找或者生成所需要的wrapper，集成计划，信息发送计划。
- 系统通过符合用户需求的wrapper或者service得到包含用户目标数据的一系列XML文档。然后集成引擎通过查询这些XML文档得到最终所需要的结果。
- 最后，这些结果按照用户定制的发送计划发送到用户的终端。

WDIWS系统体系结构

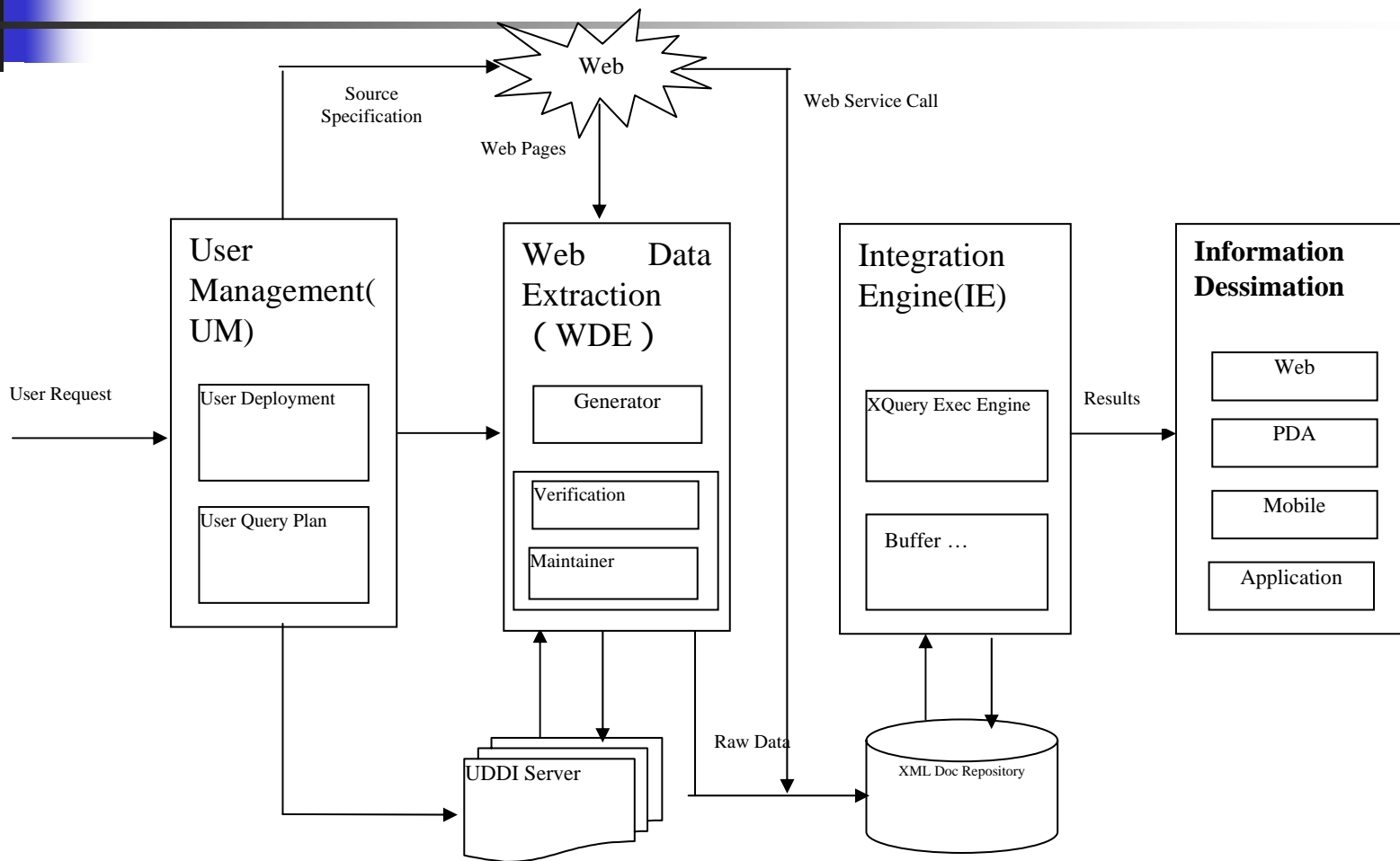


图7：WDIWS系统体系结构



WDIWS系统体系结构

- User Management(UM)
 - 系统与用户所有交互都在UM中完成。
 - 在开始抽取数据之前，每个用户都必须先指定数据源，也就是data source specification。
 - 在WDIWS系统中接受下列三种数据源
 - 来自提供Web Service服务的Web站点的数据，比如google，amazon等等
 - 可以从Web站点搜索引擎中得到的网页，比如一般的电子商务网站大多属于这种情况
 - 可以从链接结构得到的网页，比如新闻站点；这种情况下用户需要提供包含list的页面的URL，比如yahoo news的包含新闻列表的页面



WDIWS系统体系结构

- Web Data Extraction (WDE)
 - WDE工作是建立在SG-WRAP和SG-WRAM的基础之上的。这一模块的目的是从已知的一系列页面中提取出用户感兴趣的内容。
 - 在WDIWS系统中以Web 站点为单位来管理wrapper。在Web 站点内部，每一个类别（category）都有相应的wrapper。比如说，在amazon上总共有10个类别（对应于系统内部分类），这样系统就为amazon维护10个相应的wrapper。
 - 每一个已经存在的wrapper都被包装为一个service，并且在UDDI server中进行注册，使得wrapper更容易查找和管理；



WDIWS系统体系结构

- Integration Engine(IE)
 - WDIWS系统使用XQuery来表达集成计划（用户的要求都通过对XML文档的查询表达出来）。
 - 因此Integration Engine部分其实就是一个XQuery的查询引擎。
 - 其数据源是由wrapper或者网站提供的service得到的包含用户目标数据的XML文档。
 - 可以通过一个XML数据库（比如中国人民大学开发的OrientX）来实现。



报告内容

- 引言
- 信息集成技术的发展
- Web数据集成方法
- Wrapper生成技术
- Wrapper维护问题
- 基于 Web Service 的 Web 数据集成系统
(WDIWS)
- 我们的工作及展望



Our researches

■ Web Data Integration

- Schema Guided Wrapper Generation
 - SGWRAP: [ICDE2002](#), [JCST Vol.17\(4\)](#)
- Schema Guided Wrapper Maintenance
 - SGWRAM: [ICDE2003](#),
- Web Service based Data Integration

■ Native XML Database System

- OrientX: [APWeb2003](#), [软件学报](#)
- Schema based storage strategy: [VLDB2003](#)

■ Mobile Database

- Mobile Transaction: [DASFAA2001](#), [JCST Vol.17\(4\)](#), [计算机学报](#), [软件学报](#), [研究与发展](#)
- Moving Object Model - FTMOD: [DEXA2003](#)
- Moving Object Index Update: [DASFAA'2003](#)



Future Work

- 尽管目前已经有很多的集成方法和系统被提出，但存在许多不足之处。主要表现在：
 - （1）现有集成方法对Web信息的动态可变性支持不够，因此造成相应的集成系统适应性差，数据源稍有变换，即造成集成程序失效；
 - （2）集成方法的易用性差，表现在集成规则的正则表达式的表达复杂，特殊的数据结构（如数据引用）的表达也很复杂，使集成只能为一些非常熟悉系统的专业人士构造；
 - （3）信息集成的方法缺乏目的性，一般是为了集成而集成，没有与用户需求（如信息的发布）真正联系起来，使得集成系统不适合做后续的开发；



Future Work

- 针对这些现有方法存在的问题，我们提出了基于Web Services的Web数据集成技术研究方案。
 - 对于易用性差，适应性差的问题，我们拟采用基于预定义模式的数据抽取和维护方法，简化包装器生成过程，提高包装器对动态页面变化的自适应能力[19,20,21]；
 - 对于数据集成接口问题，我们将研究数据源的描述和发布规范，提供一套有效的数据源收集和选择方式，使经包装的数据源能方便的用到不同的信息集成应用中去。
 - 我们将在已有的研究成果基础上，为Web上的各类数据集成应用提供高效的开发工具和中间件。
 - 系统正在开发中.....



Copyright:

Some slides belong to:

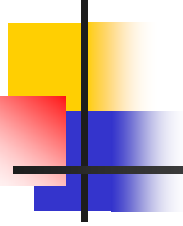
Jeffrey D. *Ullman*, *Stanford U*,
DASFAA'2003

S. Abiteboul, *INRIA and Xyleme*, WISE'2002



参考文献

- [1] S. Abiteboul, P. Buneman, and D. Suciu. Data on the Web-From Relations to Semi-Structured Data and XML, *Morgan Kauffmann Publishers*, 2000.
- [2] A.Y. Levy, A. Rajaraman, and J. J. Ordille. Querying Heterogeneous Information Sources Using Source Descriptions, In *Proc. of the Very Large Data Bases (VLDB)*, pages 251-262 Bombay, India, September 1996.
- [3] R. Bayardo, W. Bohrer, R. Brice, et al. InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments. In *Proc. of ACM SIGMOD International Conference Management of Data*, pages 195-206, Tucson, USA, May 1997.
- [4] S. Bressan, C. Goh, K. Fynn, et al. The Context Interchange Mediator Prototype. In *the ACM SIGMOD/PODS Joint Conference*, pages 525-527, Tucson, USA, May 1997.
- [5] V. Christophides, S. Cluet, J. Siméon. On Wrapping Query Languages and Efficient XML Integration. In *Proc. of ACM SIGMOD Conference on Management of Data*, pages 141-152, Dallas, USA, May 2000.
- [6] S. Cluet, C. Delobel, J. Siméon, and K. Smaga. Your Mediators Need Data Conversion! In *Proc. of ACM SIGMOD Conference on Management of Data*, pages 177-188, Seattle, USA, June 1998.
- [7] D. Chamberlin J.Robie D.Florescu. Quilt: An XML Language for Heterogeneous Data Sources. In *International Workshop on the Web and Databases (WebDB)*, Dallas, USA, May 2000.
- [8] H. Garcia-Molina, Y. Papakonstantinou, et al. The TSIMMIS Approach to Mediation: Data Models and Languages, In *Journal of Intelligent Information Systems*, 8(2): 117-132, 1997.
- [9] D. Florescu, A. Levy, and A. Mendelzon. Database Techniques for the World Wide Web: A Survey. *ACM SIGMOD Record*, 27(3): 59-74, September 1998.
- [10] I. Muslea, Extraction patterns for information extraction tasks: A survey. In *Proc. of the AAAI Workshop on Machine Learning for Information Extraction*, pages 1-6, Orlando, USA, July 1999.
- [11] B. Chidlovskii. Automatic repairing of Web Wrappers. In *3rd International Workshop on Web Information and Data Management*, pages 24-30, Atlanta, USA, November 2001.
- [12] R. Baumgartner, S. Flesca, G. Gottlob. Visual Web information extraction with Lixto. In *Proc. of the Very Large Data Bases (VLDB)*, pages 119-128, Roma, Italy, September 2001
- [13] A. Sahuguet and F. Azavant. Building Light-Weight Wrappers for Legacy Web Data-Sources Using W4F. In *Proc. of the Very Large Data Bases (VLDB)*, pages 738-741, 1999.
- [14] J. Hammer, M. Brenning, H. Garcia-Molina, et al. Template-based wrappers in the TSIMMIS system. In *Proc. of ACM SIGMOD Conference*, pages 532-535, Tucson, USA, May 1997.
- [15] I. Muslea, S. Minton, and C. A. Knoblock. STALKER: Learning extraction rules for semistructured Web-based information sources. In *Proc. of AAAI Workshop on AI and Information Integration*, pages 74-81, Madison, USA, July 1998.
- [16] C.-N. Hsu and M.-T. Dung. Generating finite-state transducers for semi-structured data extraction from the Web. *Information Systems*, 23(8): 521-538, 1998.
- [17] N. Kushmerick, D. Weil, and R. Doorenbos. Wrapper induction for information extraction. In *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, pages 729-735, Nagoya, Japan, August 1997.
- [18] L. Liu, C. Pu, W. Han. XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources. In *Proc. of International Conference on Data Engineering (ICDE)*, pages 611-621, San Diego, USA, February 2000.
- [19] X. F. Meng, H. J. Lu, H. Y. Wang, M. Z. Gu. SG-WRAP: A Schema-Guided Wrapper Generator. In *Proc. of International Conference on Data Engineering (ICDE)*, pages 331-332, San Jose, USA, February 2002.
- [20] X. F. Meng, H. J. Lu, H. Y. Wang, M. Z. Gu. Schema-Guided Data Extraction from the Web. *Journal of Computer Science and Technology(JCST)*, 17(4), 2002.
- [21] X F Meng, H Y Wang, D D Hu, M Z Gu, SG-WRAM Schema Guided Wrapper Maintenance: A Demonstration, Proceedings of ICDE2003, Mar 5-8, 2003, Bangalore, India.
- [22] N. Kushmerick. Wrapper verification. *World Wide Web Journal*, 2000, 3(2): 79-94.
- [23] N. Kushmerick. Regression testing for wrapper maintenance. In Proceedings of AAAI 1999, page 74-79, 1999.
- [24] C A Knoblock, K Lerman, S Minton, I Muslea. Accurately and Reliably Extracting Data from the Web: A Machine Learning Approach. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 2000, 23(4): 33-41.



Thank you.