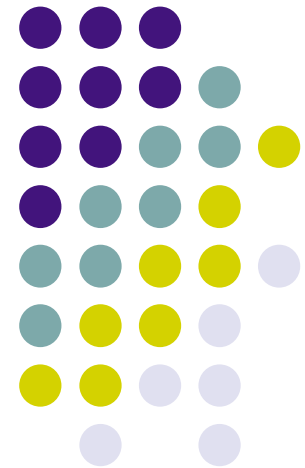


THALIA - Integration Benchmark

Mike Stonebraker
MIT

April 6, 2005



Why an Integration Benchmark?

- Information (schema) integration remains a very challenging problem
 - Significant barrier to adoption of collaborative environments
 - Termed “achilles heel” of computing [Lowell Report]
- Significant test data needed for meaningful experiments and evaluations of existing techniques
- Existing testbeds lack specific challenge queries, classification of heterogeneities, and scoring functions
 - E.g., *UIUC Web Integration Repository*, Univ. of Amsterdam’s *Web-Based Knowledge Representation Repositories* (collection of ontologies)

Overview of THALIA Testbed

1. *45 data sources* (& schemas) representing University course catalogs from CS departments worldwide
 - Well-formed, valid XML
 2. *12 challenge queries*
 - Systematic classification of the types of syntactic and semantic heterogeneities
 3. *Scoring function* for ranking the performance of an integration system
- Source data & schemas extracted with Telegraph Screen Scraper (TESS) from Berkeley
 - Wrappers preserve structural and semantic heterogeneities that exist among the different catalogs

Sample University Catalog

THALIA : **T**est **H**arness for the **A**ssessment of **L**egacy **I**nformation **I**ntegration **A**pproaches

[Home](#)
[How To Use The Benchmark](#)
[Publications](#)
[University Course Catalogs](#)
[Browse Data and Schema](#)
[Run Benchmark](#)
[Provide Feedback](#)
[Upload Your Scores](#)
[Honor Roll](#)
[People](#)
[Disclaimer & Contact](#)

Select a University to Browse its Course Catalog

Brown University
Computer Science

Home

[About Us](#)
[Contacts](#)
[News](#)
[Events/Talks](#)
[Courses](#)
• [Schedule](#)
[UTA Program](#)
[Graduate Study](#)
[Undergraduate Study](#)
[People](#)
[Research](#)
[Software Catalog](#)
[Rooms & Labs](#)
[Computing Systems](#)
[Publications](#)
[Industrial Relations](#)

Current Class Schedule

Semester 2, 2004

Course	Instructor	Title/Time	Room
CS002	Stanford	Concepts & Challenges of CS C hr. MWF 10-11	Salomon 001
CS004	Usas	Intro to Scientific Computing K hr. T,Th 2:30-4	MacMillan 117
CS016	Tamassia	Intro to Algorithms & Data Structures D hr. MWF 11-12	CIT Lubrano
CS018	Klein	CS: An Integrated Approach J hr. T,Th 1-2:30	CIT 227
CS022	Lysyanskaya	Intro. to Discrete Mathematics B hr. MWF 9-10	CIT 165
CS032	Reiss	Intro. to Software Engineering K hr. T,Th 2:30-4	CIT 165, Labs in Sunlab
CS034	Manos Renieris	Intro. to Systems Programming THURSDAY ONLY 1-2:30	TBA
CS92	Blumberg	Educational Software Seminar K hr. T,Th 2:30-4	CIT 506
CS138	Cetintemel	Networked Information Systems I hr. T,Th 10:30-12	CIT 368

Done

Extracted Source & Schema

THALIA : Test Harness for the Assessment of Legacy Information Integration Approaches

[Home](#)
[How To Use The Benchmark](#)
[Publications](#)
[University Course Catalogs](#)
[Browse Data and Schema](#)
[Run Benchmark](#)
[Provide Feedback](#)
[Upload Your Scores](#)
[Honor Roll](#)
[People](#)
[Disclaimer & Contact](#)

Select a University to Browse its XML Data and Schema

```
- <brown>
- <Course>
  <Code>CS002</Code>
  <Instructor>http://www.cs.brown.edu/~dls/>Stanford</Instructor>
  - <Title>
    http://www.cs.brown.edu/courses/cs002/>Concepts Challenges of CSC hr. MWF 10-11
  </Title>
  <Room>Salomon 001</Room>
</Course>
- <Course>
  <Code>CS004</Code>
  <Instructor>http://www.cs.brown.edu/~ausas/>Usas</Instructor>
  - <Title>
    http://www.cs.brown.edu/courses/cs004/>Intro to Scientific ComputingK hr. T,Th 2:30-4
  </Title>
  <Room>MacMillan 117</Room>
</Course>
+ <Course></Course>
+ <Course></Course>
+ <Course></Course>
+ <Course></Course>
+ <Course></Course>
+ <Course></Course>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema elementFormDefault="qualified" attributeFormDefault="unqualified" xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xs:element name="brown">
    <xs:annotation>
      <xs:documentation>Brown University</xs:documentation>
    </xs:annotation>
    <xs:complexType>
      <xs:choice minOccurs="0" maxOccurs="unbounded">
        <xs:element name="Course" minOccurs="0" maxOccurs="unbounded">
          <xs:complexType>
            <xs:sequence minOccurs="0" maxOccurs="unbounded">
              <xs:element name="Code" type="xs:string" minOccurs="0"/>
              <xs:element name="Instructor" type="xs:string" minOccurs="0"/>
              <xs:element name="Title" type="xs:string" minOccurs="0"/>
              <xs:element name="Room" type="xs:string" minOccurs="0"/>
            </xs:sequence>
            <xs:attribute name="Code" type="xs:string" use="optional"/>
          </xs:complexType>
        </xs:element>
      </xs:choice>
    </xs:complexType>
  </xs:element>
</xs:schema>
```

Integration Benchmark

- Each challenge query must be answered against at least two schemas from the testbed:
 - *Reference schema*, which is used to formulate the query
 - *Challenge schema* which exhibits the type of heterogeneity that is to be resolved
- Evaluation is based on
 - Number of correctly answered queries
 - Amount of programmatic “integration effort”

Challenge Queries

- Written in XQuery v. 1.0.
- Each query illustrates different type of heterogeneity
- *Attribute heterogeneities* - 5 queries
 - Exist between two related attributes in different schemas
- *Missing data* - 3 queries
 - Due to missing information (structure or value) in one of the schemas
- *Structural heterogeneities* - 4 queries
 - Due to discrepancies in the way related information is modeled/represented in different schemas
- In each group, heterogeneities are organized in increasing order of complexity of the effort that is needed to resolve them

Sample Challenge Queries

- *Attribute heterogeneities*
 - Find all database courses that meet at 1:30pm on any given day
 - **Challenge:** Conversion of time represented in 12 hour-clock to 24 hour-clock
 - **Challenge:** Non English language schemas
- *Missing data*
 - List all textbooks for courses about verification theory
 - **Challenge:** Proper treatment of NULL values (e.g., no textbook value attribute in challenge schema)
- *Structural heterogeneities*
 - Find the room in which the database course is held
 - **Challenge:** Room information in different places in reference and challenge schema

Plans and Future Work

- Adding additional data sources
 - Provide additional sources of heterogeneity
 - New challenge queries
- Soliciting feedback on the usefulness of the benchmark
 - Results of running our benchmark on as many of the existing integration systems as possible

Web Site

<http://www.cise.ufl.edu/project/thalia.html>