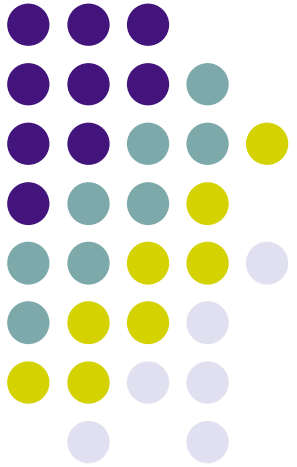
THALIA - Integration Benchmark

Mike Stonebraker

MIT



April 6, 2005

Why an Integration Benchmark?

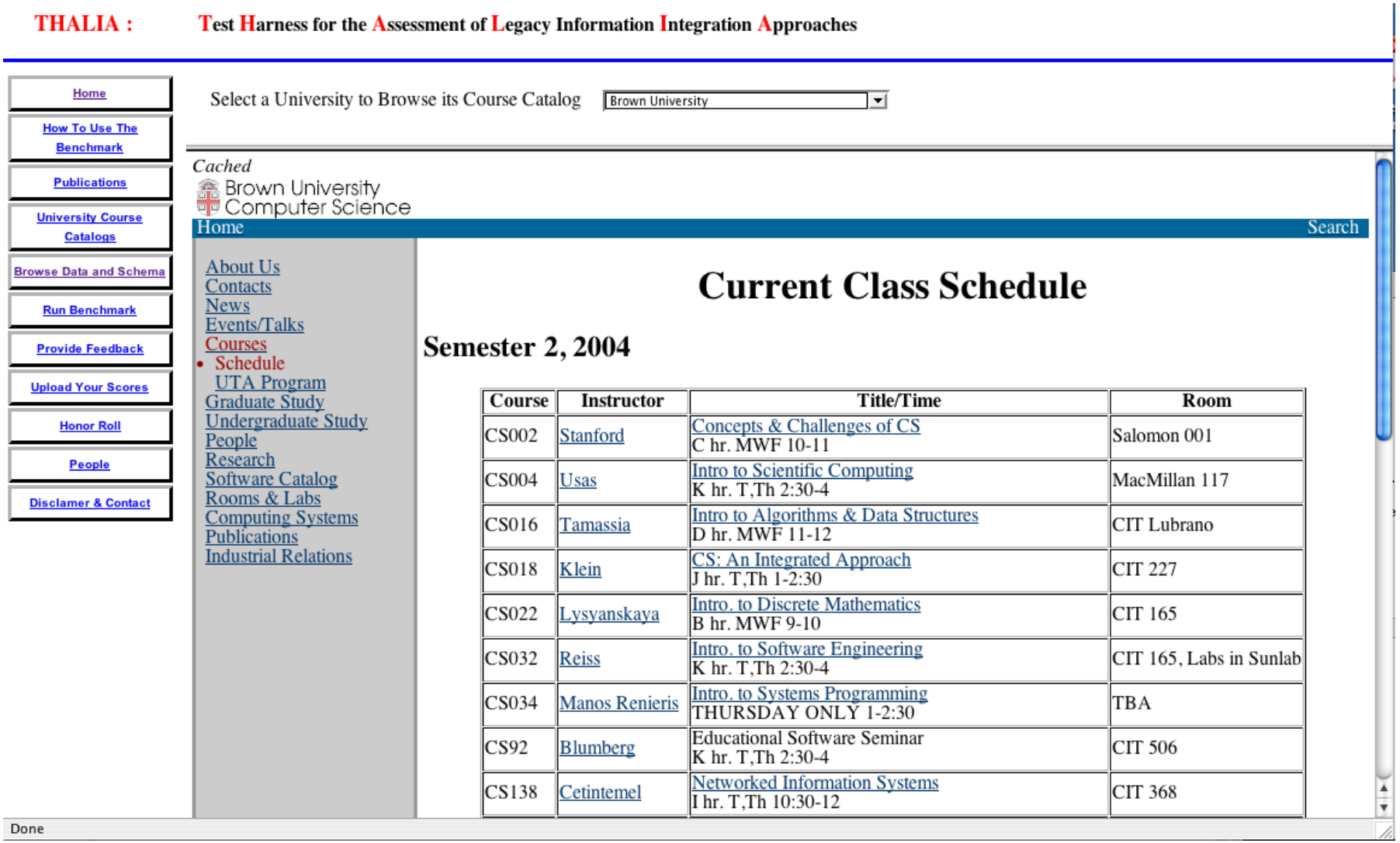
- Information (schema) integration remains a very challenging problem
  - Significant barrier to adoption of collaborative environments
  - Termed “achilles heel” of computing [Lowell Report]
- Significant test data needed for meaningful experiments and evaluations of existing techniques
- Existing testbeds 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

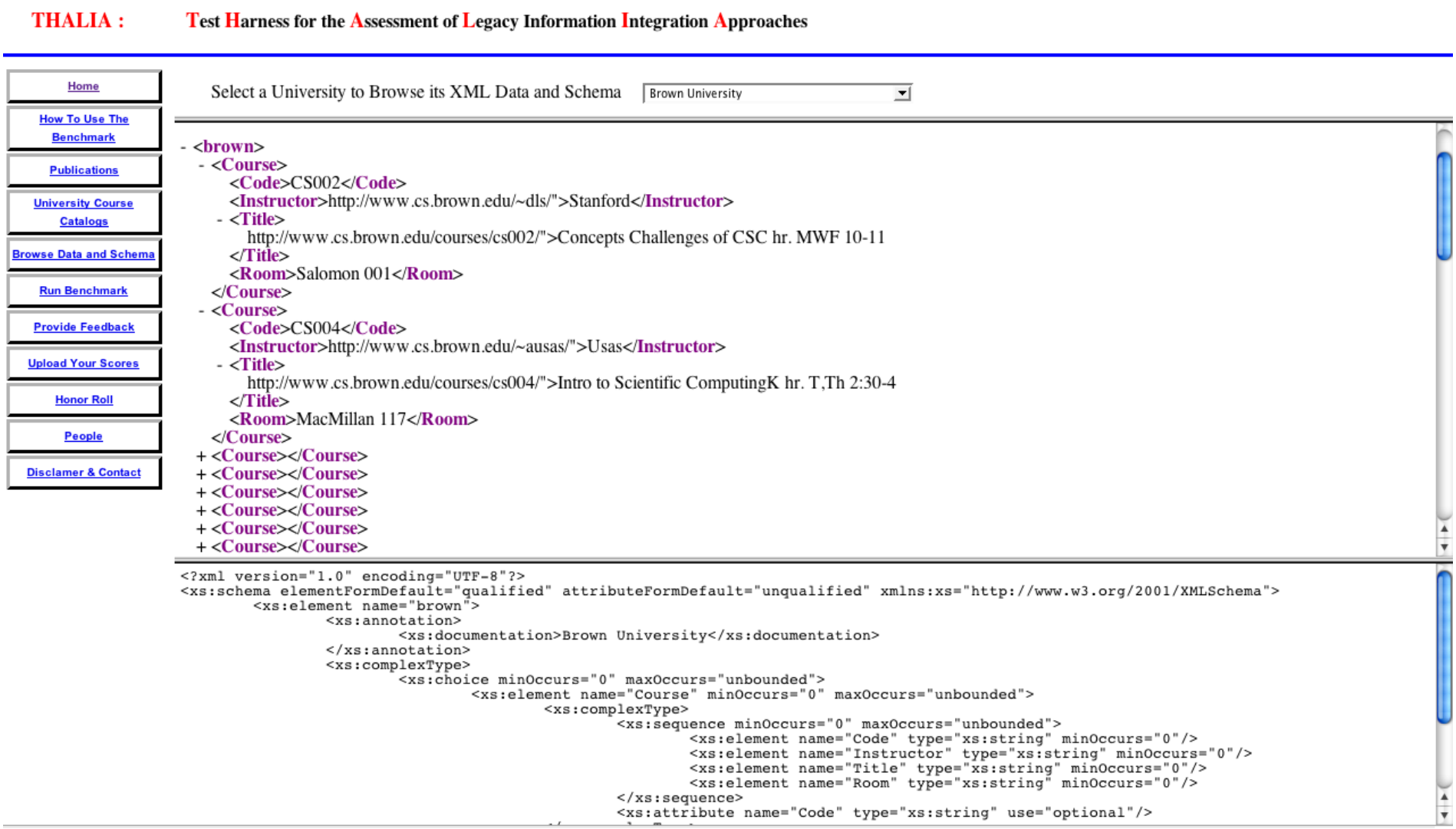
1. *45 data sources* (& schemas) representing University course catalogs from CS departments worldwide
   - Well-formed, valid XML
2. *12 challenge queries*
   - Systematic classification of the types of syntactic and semantic heterogeneities
3. *Scoring function* for ranking the performance of an integration 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



Extracted Source & 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