

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

Joachim Hammer
Computer & Information
Science & Engineering
University of Florida
jhammer@cise.ufl.edu

Mike Stonebraker
Computer Science & Artificial
Intelligence Laboratory
MIT
stonebraker@lcs.mit.edu

Oguzhan Topsakal
Computer & Information Science
& Engineering
University of Florida
otopsaka@cise.ufl.edu

1. Introduction

We introduce our new, publicly available *testbed and benchmark called THALIA*¹ (Test Harness for the Assessment of Legacy information Integration Approaches) for testing and evaluating integration technologies. THALIA (see: <http://www.cise.ufl.edu/research/dbintegrate/>) provides researchers with a collection of 40 downloadable data sources representing University course catalogs from computer science departments worldwide. In addition, THALIA currently provides a set of *twelve challenge queries* as well as a scoring function for ranking the performance of an integration system. A second contribution is a *systematic classification of the types of syntactic and semantic heterogeneities*, which directly lead to the twelve challenge. We have chosen course information as our domain of discourse because it is well known and easy to understand. Furthermore, there is an abundance of data sources publicly available that allowed us to develop a testbed exhibiting all of the syntactic and semantic heterogeneities that we have identified.

We believe that THALIA will not only simplify the evaluation of existing integration technologies but also help researchers improve the accuracy and quality of future approaches by enabling more thorough and focused testing.

While testbeds for evaluation of integration technologies are not new (see, e.g., the *Web-Based Knowledge Representation Repositories* project at the Vrije Universiteit Amsterdam at <http://wbkr.cs.vu.nl/> or the *UIUC Web Integration Repository* at <http://metaquerier.cs.uiuc.edu/repository>), what distinguishes THALIA is the fact that it combines rich test data with a set of queries and associated scoring function to enable the objective evaluation and comparison of integration systems.

2. Testbed

In THALIA, University course catalogs are represented using well-formed and valid XML according to the extracted schema for each course catalog. The schema of the XML document remains as close to the original schema of the corresponding catalog as possible. Extraction and translation from the original representation is done using a source-specific wrapper. Although the wrapper removes heterogeneities at the system-level it preserves structural and semantic heterogeneities that exist among the different catalogs.

We used a modified version of the Telegraph Screen Scraper (TESS) developed at UC Berkeley (see <http://telegraph.cs.berkeley.edu/tess/>) to extract source data and produce the XML output that makes up the testbed. We made several modifications to the original TESS source code since it was initially programmed to flatten nested substructure in the output document which would eliminate many useful heterogeneities.

3. Integration Benchmark

The Integration Benchmark consists of a set of twelve *challenge queries* that illustrate the different types of heterogeneities in the testbed. Rather than adopting the commonly-used but very general distinction between structural and semantic heterogeneities, we divided the twelve cases into three groups, based on commonalities among the different heterogeneities: (1) *Attribute heterogeneities* (5 queries), which exist between two related attributes in different schemas; (2) *missing data* (3 queries), which are due to missing information (structure or value) in one of the schemas; and (3) *structural heterogeneities* (4 queries), which are due to discrepancies in the way related information is modeled/represented in different schemas. In each of the three groups, heterogeneities are organized in increasing order of complexity of the effort that is needed to resolve them. Our list of twelve cases is not an exhaustive list as different application domains may exhibit additional

¹ Derived from Greek *thallein* meaning "to blossom".

heterogeneities not seen in the current collection of course catalogs. However, the heterogeneities exhibited in our testbed cover most of the heterogeneities discussed in the literature (e.g., [2, 3]) and thus represent a reasonable subset of cases that must be resolved by an integration system.

Each benchmark query must be answered against at least two schemas from the testbed: a *reference schema*, which is used to formulate the query, as well as a *challenge schema* which exhibits the type of heterogeneity that is to be resolved by the integration system. In some cases a query may illustrate additional types of heterogeneities that are also showcased in other queries. Queries are written in XQuery v. 1.0.

Using THALIA, integration systems can be evaluated and compared based on the number of correctly answered benchmark queries as well as on the amount of programmatic “integration effort” measured in terms of the size of the integration specification that must be invested in order to answer each query.

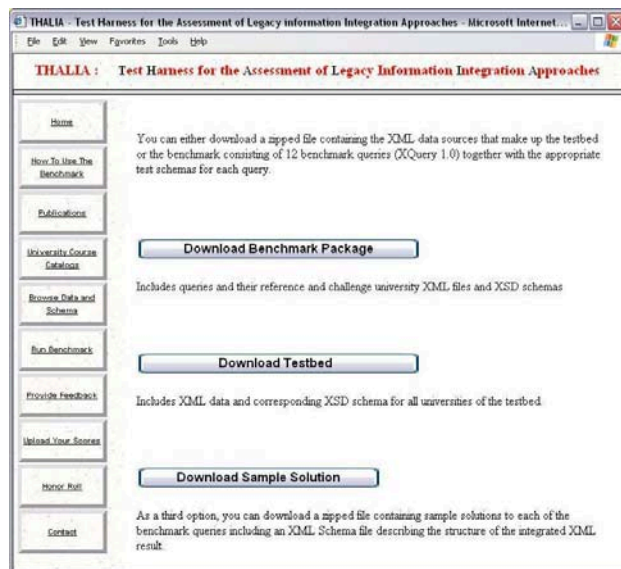


Figure 1. Snapshot of THALIA's Web Site displaying three different “Run Benchmark” options

4. Web Site

We have created a Web site to simplify the access to the testbed and benchmark. A snapshot of the home page is shown in Figure 1 depicting the available interface options in the left-hand frame. Specifically, THALIA supports the browsing of the available course catalogs in their original representation (‘University Course Catalogs’ button), as well as the viewing of the extracted XML documents and corresponding schemas (‘Browse Data and Schema’).

The ‘Run Benchmark’ option provides the user with three choices: (1) Download a zipped file containing the XML and XML Schema files of all available course catalogs. We believe this option will be useful to those researchers and developers looking for test data for the experimentation with new integration solutions. (2) Download a zipped file containing the twelve benchmark queries as well as the corresponding test data sources. We believe this option will be useful to researchers wishing to evaluate the capabilities of their integration solution using a third-party, independent benchmark. Furthermore, the benchmark scores will be the first publicly available measure of integration technologies allowing the objective comparison of integration solutions. Note that integration systems that do not provide query processing can still use the benchmark by providing an integrated schema over the two data sources associated with each benchmark query. (3) Download a zipped file containing sample solutions to each of the benchmark queries including an XML Schema file describing the structure of the integrated XML result.

In order to facilitate the comparison of technologies, we invite users of the benchmark to upload their benchmark scores (‘Upload Your Scores’) which can be viewed by anybody using the ‘Honor Roll’ button.

5. Conclusion and Proposed Directions

THALIA and its related efforts have been a direct response to the Lowell Report [1], which asks for exactly this sort of scheme to stimulate further research in real world integration problems. As preliminary results have shown, current systems do not score well on our benchmark, and we hope that THALIA will be an inducement for research groups to construct better solutions.

We are currently adding new data sources to the testbed, which will provide additional sources of heterogeneity. We are also soliciting feedback on the usefulness of the benchmark including the results of running our benchmark on as many of the existing integration systems as possible.

References

- [1] J. Gray, H. Schek, M. Stonebraker, and J. Ullman, "The lowell report," *Proc. 2003 ACM SIGMOD International Conference on Management of Data*, San Diego, CA, 2003.
- [2] S. Navathe, R. Elmasri, and J. Larson. Integrating user views in database design. *Computer*, **19**(1):50-62, 1986.
- [3] A. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Computing Surveys*, **22**(3):183-236, 1990.