

Brainstorming Session

Tuesday, Oct 2, 11:30 – 1:00

Phil Bernstein, Chair

We started by spending 10 minutes listing for possible discussion topics:

1. Usage scenarios (in addition to data exchange and OR mapping)
 - a. How to reduce the skill level to enable users?
2. Theory results being ignored by systems people
3. What should a knowledge base consist of?
 - a. How to create and reuse knowledge?
 - b. How to create a community resource that people will use?
4. How can we create a testbed or benchmark?
 - a. How can we compare systems?
5. Is metadata repository technology sufficient for the problems that we're interested in?

We voted and decided to go after issue 4.

How can we compare systems?

How can we create a testbed or benchmark?

Lack of an end-to-end testbed of tools (Peter Mork @ MITRE, MetaMatrix is going open source, Sergey Melnik's Rondo):

- Schema import
- GUI tool
- Matcher
- Mapping generation
- Query processor and/or data exchange engine

Philadelphia Workshop suggested developing a standard mapping specification language. <http://db.cis.upenn.edu/iiworkshop/>

Thalia benchmark at University of Florida, in support of Morpheus. 60 XML schemas, but concepts are just courses and departments. Challenge queries.

OAEI has a metamodel for correspondences ("alignment format"). Was funded by EU project: Knowledge Web.

We need both micro-level and macro-level benchmarks.

Would a solution to scientific data integration be suitable for commercial data integration? In any case, the former is worth research investigation.

How much industrial artifacts can be made available in Open Source environment?

Talend – open source tool for ETL. <http://www.talend.com/ressources/index.php>

Recipe for comparing systems:

1. Benchmark
2. Metrics – full cost-of-ownership
3. Methodology for comparing systems

What is the benefit of doing a data integration project to an organization?

We ran out of things to say about this topic, so we switched to issue 3.

Knowledge Base within a system

How to reuse information within an enterprise?

Set it up as a community who share this reusable information.

Community development of an ontology and its mapping to custom schemas. E.g., see Schema Logic.

Semantic Web has a taxonomy of subjects. Papers are tagged so they can be clustered, social networks can be analyzed, etc.

There's a history of projects on reuse repositories for schemas.

Need a tool to create, manage and use controlled vocabulary.

We ended with a few minutes about issue 2.

Theory

Had a brief discussion about looking at long-term deep questions that are ahead of their time, versus identifying theoretical questions of direct interest to today's systems.