Making data integration research transferable to product

Arnon Rosenthal
The MITRE Corporation

Data integration has been a database and AI research topic for decades, but industrial strength integration systems embody very few of these research results. The DILS community may benefit from understanding the transition barriers experienced by mainstream integration (XML and relational structures), and possible mitigation strategies.

For the business market, commercial tool suites (at >$200K!) provide GUIs for manual specifications, glue code (e.g., transformations and wrappers), data profilers, rule languages, and code generators. (Query processors, a separate market, will not be addressed here). Database and AI research offer great promise for automated assistance and automation, to reduce the time and skills required of human integrator. We address four transition barriers.

- Much research from each viewpoint (ontology, DB) ignores the other community's achievements and concerns (e.g., standard logics, tool frameworks, attracting purchasers). We describe opportunities for cross-fertilization.
- Researchers often focus on automating "upstream" tasks. Consider schema integration / ontology alignment tools, whose output describes attribute correspondences. A programmer must manually disambiguate (e.g., should join preserve unmatched items?), insert transformations, and weave them into relational/XML queries that end users can run. Each subsequent schema change will again require a programmer downstream. The work process is complex, with modest benefit. Would your programmers buy and use such a tool? *Downstream principle: automate the last remaining manual step before the end user*.
- A researcher's bright idea needs to be added to a system that has critical mass. Today, each vendor produces (slowly, expensively) a tightly coupled suite. Open source (e.g. Red Hat MetaMatrix, or MITRE's Harmony) would speed the incorporation of capabilities outsiders develop, notably those for life sciences. It could reduce the need for government-to fund life sciences suites. They may also reduce barriers to market entry, and thus prices.
- Researchers rightly present results that require preconditions, e.g., handling only target schemas that lack constraints or are acyclic. But then they stop, providing no guidance about messy real systems that do not satisfy those preconditions. Researchers should decompose the *general* problem, using their technique on some parts and leaving a residue simpler than the original task. For example, if a target schema's constraints make full alignment intractable, align to the target structure with an easier constraint set. The residue (consistent schemas, different constraints) is now a simpler and smaller task.

ml 19-10-07 14:26
Eliminato: 8/24/2007

ml 19-10-07 14:26
Eliminato: 3:48:36 PM

Arnon Rosenthal 24-8-07 15:48
Inserimento: 8/24/2007

ml 19-10-07 14:26
Eliminato: 5/16/2007

Arnon Rosenthal 24-8-07 15:48
Inserimento: 3:48:36 PM

ml 19-10-07 14:26
Eliminato: 12:55:58 AM

Many additional materials on pragmatics of data integration appear at http://www.mitre.org/staffpages/arnie/

Acknowledgment: The author would like to thank Len Seligman, Peter Mork, Barbara Blaustein, and Jean Stanford for suggesting many improvements.

ml 19-10-07 14:26
Eliminato: 8/24/2007

ml 19-10-07 14:26
Eliminato: 3:48:36 PM

Arnon Rosenthal 24-8-07 15:48
Inserimento: 8/24/2007

ml 19-10-07 14:26
Eliminato: 5/16/2007

Arnon Rosenthal 24-8-07 15:48
Inserimento: 3:48:36 PM

ml 19-10-07 14:26
Eliminato: 12:55:58 AM