

Data Integration as an Industrial Process

Len Seligman, Arnon Rosenthal
703-983-5975 • seligman@mitre.org
781-271-7577 • arnie@mitre.org

MSR

The logo for the MITRE Technology Program, featuring a stylized graphic of stacked blocks in yellow, orange, and blue to the left of the text.

MITRE
Technology
Program

Problem

Data Integration - the ability to *meaningfully* exchange data among separately developed information systems

- Data integration costs too much
 - You have to know too much
 - Databases
 - Application programming
 - Distributed systems programming
 - and*
 - The application domain
 - Too little reuse of expensively gathered knowledge

Vision: An “Industrial” Process

- **Modularize the process as small steps, each using one skill**
 - Easier to automate pieces
 - Can combine small point solutions
 - Move from highly skilled craftsmen to specialists
- **Incentives, not mandates, for integrators**
 - Metadata will be good if it enables providers to meet their responsibilities
 - Capture knowledge and generate code that providers want (“describe and generate”)

A. Rosenthal, L. Seligman, S. Renner, F. Manola, “Data Integration Needs an Industrial Revolution,” *Int. Workshop on Foundations of Models for Information Integration (FMII)*, Sept. 2001

Objective

- Refine the industrial process approach
- Validate hypotheses:
 - Integration tasks can be modularized
 - “Describe and generate” is achievable
 - “Describe and generate” provides benefit
- Move industry, researchers, and sponsors toward the vision
 - *If tools are ready, insert them on real projects*

Activities

- **Modularized integration steps and split skills required to perform them**
- **Described pragmatic needs not met by current community research agenda**
- **Compared (scalable) profile-driven vs. traditional federated integration approaches. Devised hybrid to combine their strengths.**
- **Constructing experimental framework and metrics to evaluate industrial approach and compare integration techniques**
- **Experimenting with “describe and generate” tools and real data**

Highlight: Integration Survey

Data Integration/Interoperability Survey

1. Please check all that apply:

- I have done research in data integration/interoperability
- I have participated in a small data integration effort (8-20 entities) with real schemas and data.
- I have participated in a large data integration effort (> 20 entities) with real schemas and data.

2. For a small data (but nontrivial) integration/interoperability project (e.g., 8-20 entities to be integrated), estimate the typical percentage of effort expended on each of the following tasks (total should equal 100):

a. Gather knowledge about sources

c. Identify semantic correspondences among sources
and from sources to consumer views

_____ %

d. Specify and create needed attribute transformations

_____ %

e. Specify data combination/data cleaning rules

_____ %

f. Create logical mappings (e.g., SQL views) from sources to consumer

_____ %

g. Create and optimize an executable connection for specific
run-time environment

_____ %

Surveying practitioners and researchers

- First attempt to determine where the time goes
- Do vendors and researchers work on the wrong issues?

Demonstration: Clio

The screenshot displays the Clio software interface, which is used for mapping data between source and target schemas. The interface is divided into three main sections: Source Schemas, Target Schema, and a data preview table.

Source Schemas: This section shows two sets of schemas. The first set is 'Set of (AIRLINETYPE)', containing 'AIRLINETYPE: Record' with fields 'AIRLINETYPE (String)' and 'TYPE (String)'. The second set is 'Set of (ETMSMAIN)', containing 'ETMSMAIN: Record' with fields: 'MSGID (Float)', 'FLIGHTID (Float)', 'MSGDATE (Time)', 'SEQNO (String)', 'GMTTIMESTAMP (Float)', 'MSGTYPE (String)', 'CARRIER (String)', 'FLIGHTNO (String)', 'COMPUTERID (String)', 'FACILITYID (String)', 'DEPTAIRPORT (String)', 'ARRAIRPORT (String)', 'DEPTTIME (String)', 'ARRTIME (String)', 'RAWDATA (String)', 'AMENDDATA (String)', 'EQUIPID (String)', 'EQUIP (String)', and 'EQUIPTYPE (String)'. Arrows indicate mappings from these fields to the target schema, with percentages such as 66.7% for MSGID and 100.0% for others.

Target Schema: This section shows three sets of schemas. The first is 'Set of (ETMSCOMMON)', containing 'ETMSCOMMON: Record' with fields: 'MSGID (Float)', 'FLIGHTID (Float)', 'MSGDATE (Time)', 'SEQNO (String)', 'GMTTIMESTAMP (Float)', 'MSGTYPE (String)', 'CARRIER (String)', 'FLIGHTNO (String)', 'COMPUTERID (String)', and 'FACILITYID (String)'. The second is 'Set of (ETMSDZ)', containing 'ETMSDZ: Record' with fields: 'MSGID (Float)', 'EQUIPINDICATOR (String)', 'EQUIP (String)', 'EQUIPTYPE (String)', 'DEPTAIRPORT (String)', 'DEPTTIME (String)', 'ARRAIRPORT (String)', and 'ARRTIME (String)'. The third is 'Set of (ETMSFZ)', containing 'ETMSFZ: Record' with field 'MSGID (Float)'.

Data Preview Table: The table on the right shows the resulting data for the 'ETMSMAIN' schema. It has two columns: 'MSGID' and 'Sk68(MSGID)'. The data rows are as follows:

| MSGID | Sk68(MSGID) |
|---------------|---------------|
| = MSGID | = Sk68(MSGID) |
| = Sk69(MSGID) | = Sk69(MSGID) |
| = null | = null |
| = null | = null |
| = null | = null |
| = null | = null |
| = null | = null |
| = null | = null |
| = null | = null |
| = null | = null |

- After 15 years of research, prototype integration toolkits are ready to be tried on real projects
 - We may be the first

Impacts

- **Influencing emerging industrial tools and researchers' agendas**
- **First serious effort to evaluate integration technologies**
 - **Providing metrics where none previously existed**
 - **Gaining insight into next-generation commercial tools**
- **Influencing sponsor strategies - e.g.,**
 - **DOD Data Interoperability Broad Area Review**
 - **“Data Integration without a Blueprint” for fast moving programs**
 - **Neuro-Informatics: Enabling use of emerging tools**

Future Plans

- **Survey practitioners to determine where costs are greatest**
- **Conduct experiments using research prototypes**
 - **Aviation, brain mapping, and tax administration data**
- **Adapt metrics to improve project planning**
- **Continue publishing and transitioning results to MITRE and sponsor projects**