Bio/Ecosystem Informatics

Renée J. Miller
University of Toronto

DB research problem:
managing data semantics
Managing Data Semantics

- Semantics modeled by
  - Schemas (structure and constraints)

- Creating semantics
  - Data design

- Managing semantics
  - Understanding and reconciling different choices made in modeling data semantics
Managing Data Semantics

- **Problem**
  - Active area of DB research
  - Not specific to bio/ecosystem informatics

- **Solutions**
  - Tailored to data characteristics
  - Our solution is focused on curated data
Clio: Schema & View Management

University of Toronto
- Renée J. Miller
- Periklis Andritsos, Ariel Fuxman,
- Tasos Kemesetsidis, Yannis Velegrakis

IBM Almaden
- Laura Haas
- Ron Fagin, Mauricio Hernández
- Howard Ho, Phokion Kolaitis
- Lucian Popa, Ling-Ling Yan
Clio: Schema & View Management

- Manage datasets that may differ in:
  - Data models
  - Structures (table or nesting structure)
  - Semantics (constraints)
  - Content

- Example problems
  - Schema Mapping
  - Schema Translation (Wrapper Generation)
  - Schema Evolution
  - Schema Integration…
Clio facilitates:
- Creation, management, debugging of **schema mappings**
- Uses
  - **Data exchange**: create a (materialized) target instance
  - **Data integration**: translate queries on (virtual) target into queries on source(s)
Data Publishing & Exchange

Microarray data in DB2

Relational Schema

XML schema

WWW

Business partner

WebService

WSDL

Life Sciences Company

SwissProt

PubMed

GenBank

...
Illustration: Clio Schema Mapping
Illustration: Clio Schema Mapping

- Support Nested Structures
- Element correspondences
  - Human friendly
  - Automatic discovery
- Preserve data meaning
  - Discover data associations
  - Use constraints & schema
- Create New Target Values
- Produce Correct Grouping
- And produce XQuery
Why Should DB Researchers Care about Bio/Ecosystems?

- Great Data
  - Rich, complex semantics
  - Exercises many facets of our models
- Publicly available!!!
- Domain curators are expert DBAs
  - Provide important feedback
How Should DB Researchers Work with Bio/Ecosystem Scientists?

- Our Model
  - Took existing, general DM problem
  - Explored systems and foundational issues
  - Prototyped tool
  - Team of developers work closely with domain scientists
  - Provide feedback

*Arms-length collaboration!!!*
Overview

Goal: interoperability between independent data sources
- Schema Mapping
- Data Translation (Data Exchange Problem)

Challenges
- Schemas can be arbitrarily different
- Still, data must not lose its meaning during translation
  - Maximum advantage of semantics embedded in schemas & data
    - Used in compilation
  - Facilitate user specification of any additional semantics
    - As by-product user learns if semantics incorrect/incomplete
    - Performed manually: complex user queries, programs, etc.

Output: correct data translation program