DataFoundry: An Approach to Scientific Data Integration

Terence Critchlow

Ron Musick **Ida Lozares**

Center for Applied Scientific Computing

Tom Slezak **Krzysztof Fidelis**

Biology and Biotechnology Research Program

Lawrence Livermore National Laboratory

IBC Bioinformatics

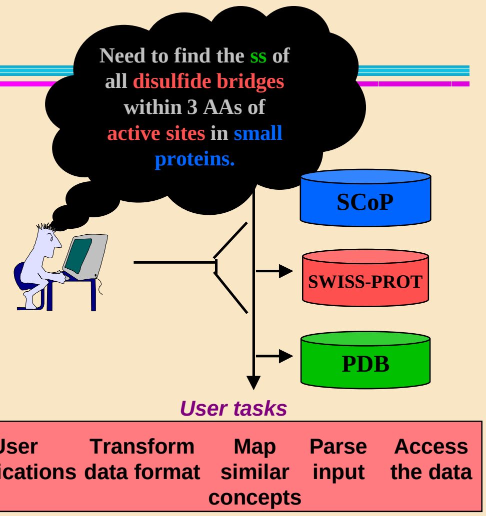
October 1, 1999

Outline

- **Motivation**
- **DataFoundry's integration strategy**
- **Improving user interfaces**
- **Beyond fully integrated data**
- **Conclusions**

Current environment

- Data is
  - Hard to find
  - Hard to understand
  - Hard to reconcile
  - Hard to analyze

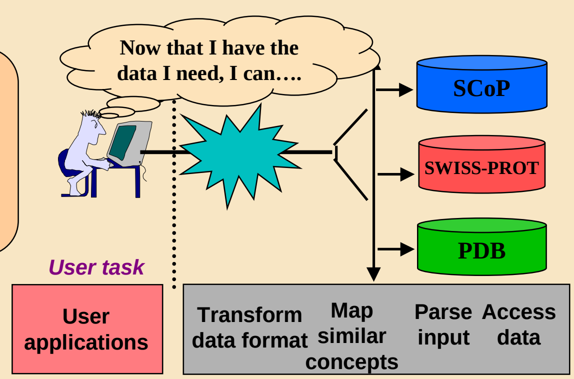


Scientists waste time and energy doing data management.

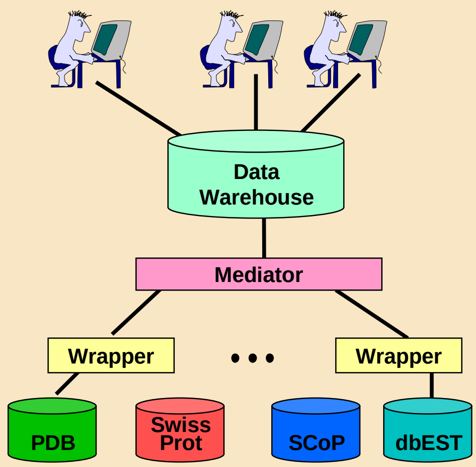
What is our ideal environment?

A *single* location that provides *effective* access to a *consistent* view of data from *many* sources through an *intuitive and useful* interface.

Businesses use data warehouses to accomplish this.

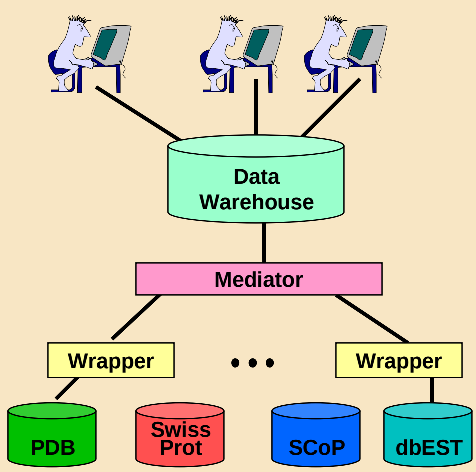


Data warehouses



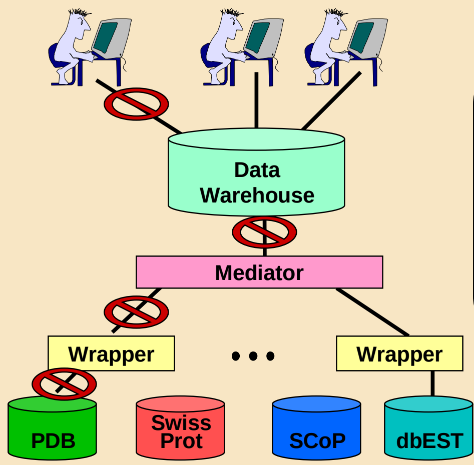
A data warehouse is a repository that provides a single access point to a collection of data obtained from a set of distributed, heterogeneous sources.

Data warehouses



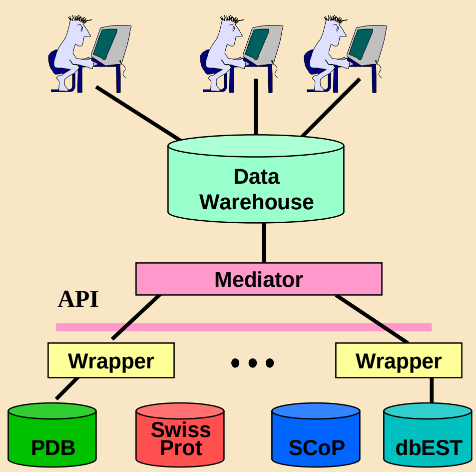
- Interfaces
  - provide intuitive access to the data
  - possibly change data format to meet user expectations
- Warehouse
  - stores a consistent view of data in a local repository
- Mediator
  - transform data from source format to warehouse format
- Wrappers
  - read data from source into internal representation

Warehouses don't work in dynamic domains



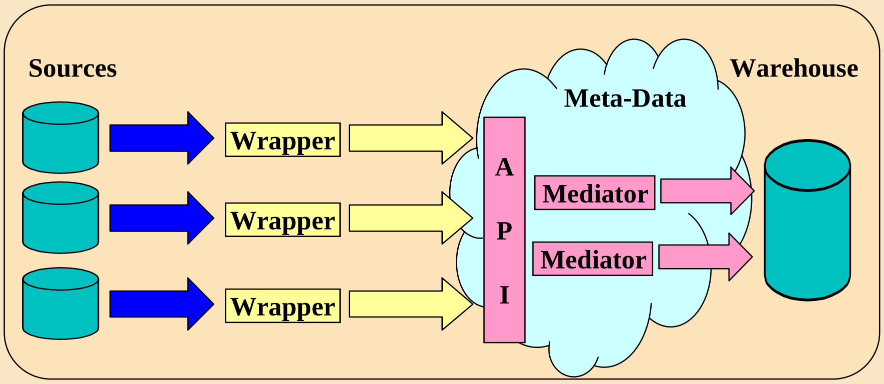
When schemata are modified,
or new sources are added,
wrappers and mediators break.

Key insight

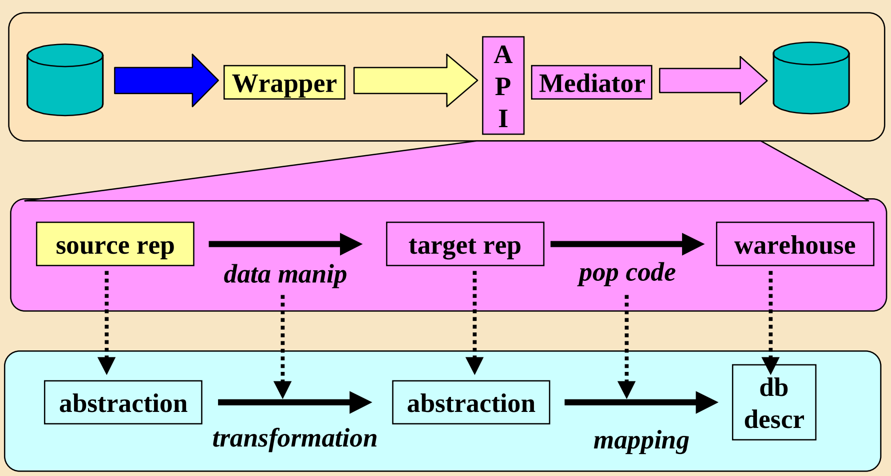


Extensive use of meta-data can dramatically reduce maintenance costs.

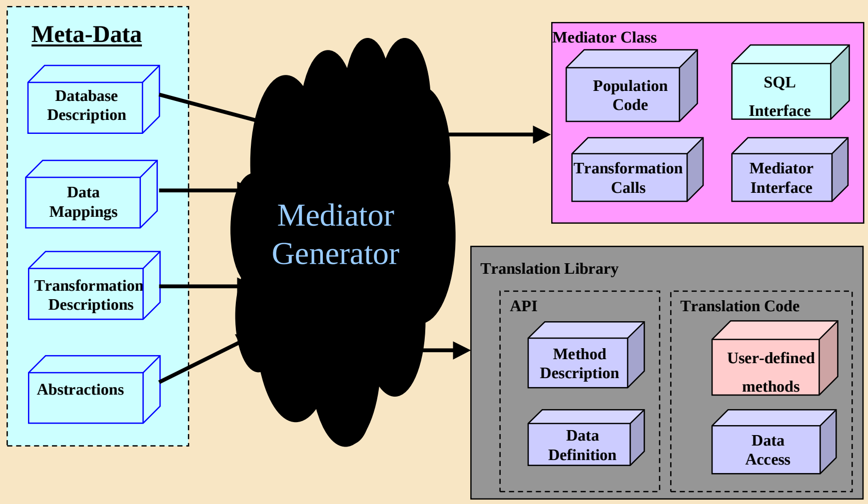
The DataFoundry approach:



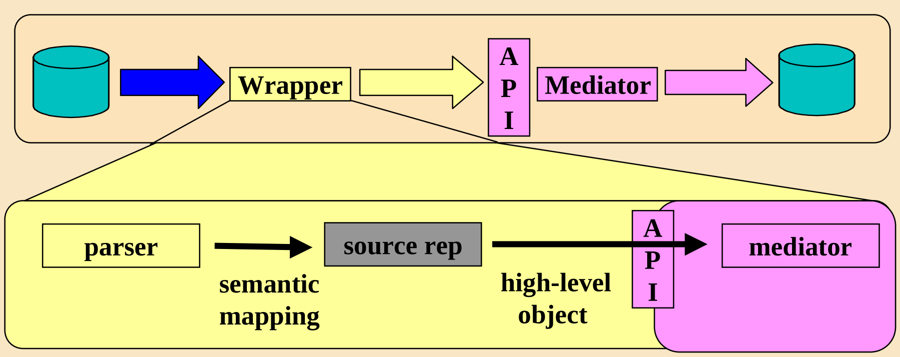
Four types of meta-data are required



Generating the mediators



The translation library and mediator class are used by the wrapper



Results:

Integrating SCoP into warehouse that already contains PDB and SWISS-PROT.

Activity/ integration style	manual	meta-data	diff	%diff
understanding SCOP	2.0	2.0	0.0	0
writing wrapper	4.5	2.5	2.0	44%
modifying schema	0.5	0.5	0.0	0
writing mediator	4.0	0.0	4.0	---
modifying meta-data	0.0	1.0	(1.0)	---
total time in days	11.0	6.0	5.0	45%

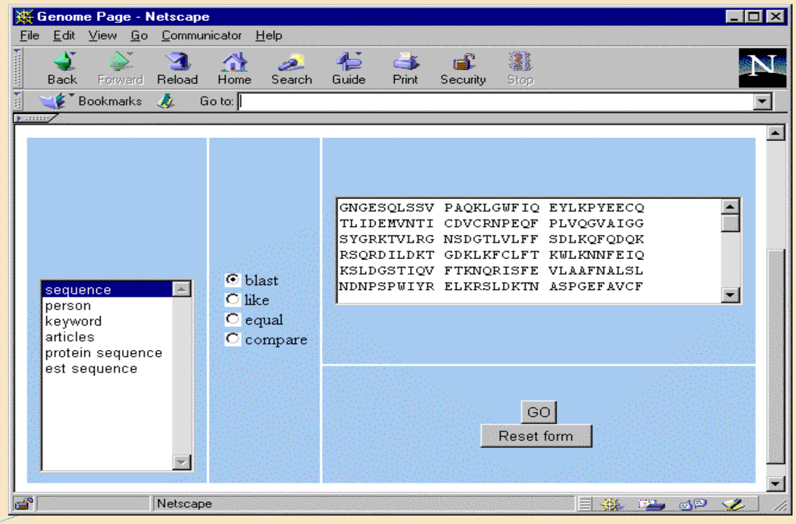
Improving Data Access

Scientists need

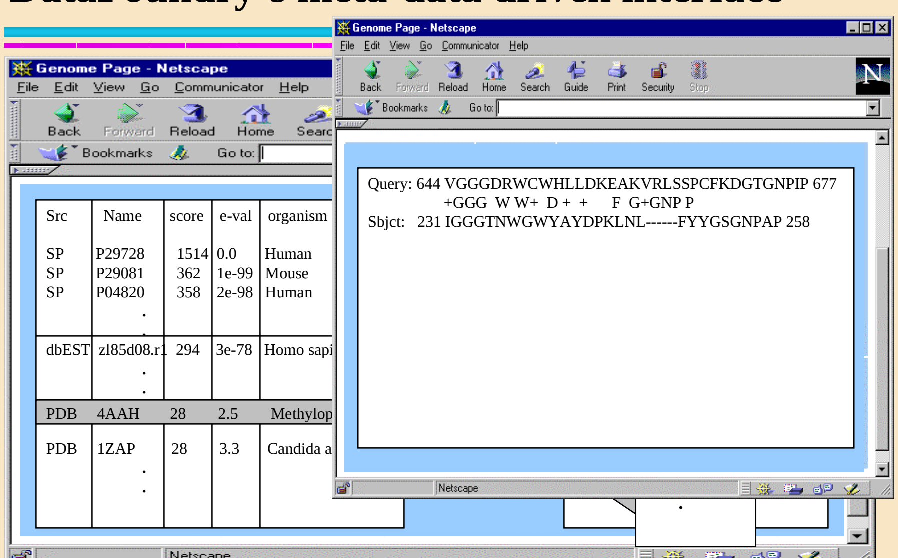
- **Better access to the data**
  - combine data from multiple sources
  - annotate data
  - perform complex queries
- **Better functionality**
  - customized notification messages
  - integrated interface to tools
  - personalized responses to queries

By extending our meta-data representation, we can provide powerful, customizable access to data.

DataFoundry’s meta-data driven interface

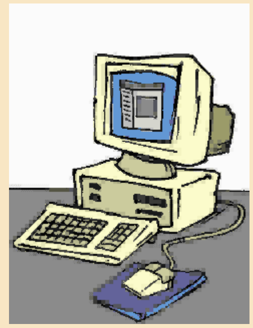


DataFoundry’s meta-data driven interface



Beyond fully integrated data

- There are over 500 genomics data sources available on the web.
- Scientists want as much *relevant* information as possible.
- Integrating data from all of these sources is impossible.



Semantically integrating critical data sources, and providing basic access to others, offers the best possible solution.

Summary

Scientists need *intuitive* access to data from both internal and external sites.

- **integrated data**
  - allows complex queries
  - is easier to understand
  - limits the number of sites
- **non-integrated data**
  - allows more sites
  - is more flexible
  - is harder to query

Conclusions

- Meta-data provides a way to
  - reduce the cost of integrating new sources
  - reduce the cost of accessing non-integrated sources
  - provide a powerful, and *intuitive*, query mechanism
  - customize the user interface

DataFoundry is building on its meta-data based infrastructure to develop a scalable, flexible, and useable system.

DataFoundry: An Approach to Scientific Data Integration

Terence Critchlow

Center for Applied Scientific Computing
Lawrence Livermore National Laboratory

critchlow@llnl.gov
www.llnl.gov/casc/datafoundry/

Work performed under the auspices of the U.S. DOE by LLNL under contract No. W-7405-ENG-48. UCRL-JC-134854