Dataspaces: The Tutorial

Alon Halevy, David Maier

VLDB 2008

Auckland, New Zealand

Outline

- Dataspaces: why? What are they?
  - Examples and motivation
- Dataspace techniques:
  - Locating and understanding data sources
  - Creating mappings and mediated schemas
  - Pay-as-you-go: improving with time
  - Query processing for dataspaces
- Research challenges on specific dataspaces:
  - Science, the desktop, the Web

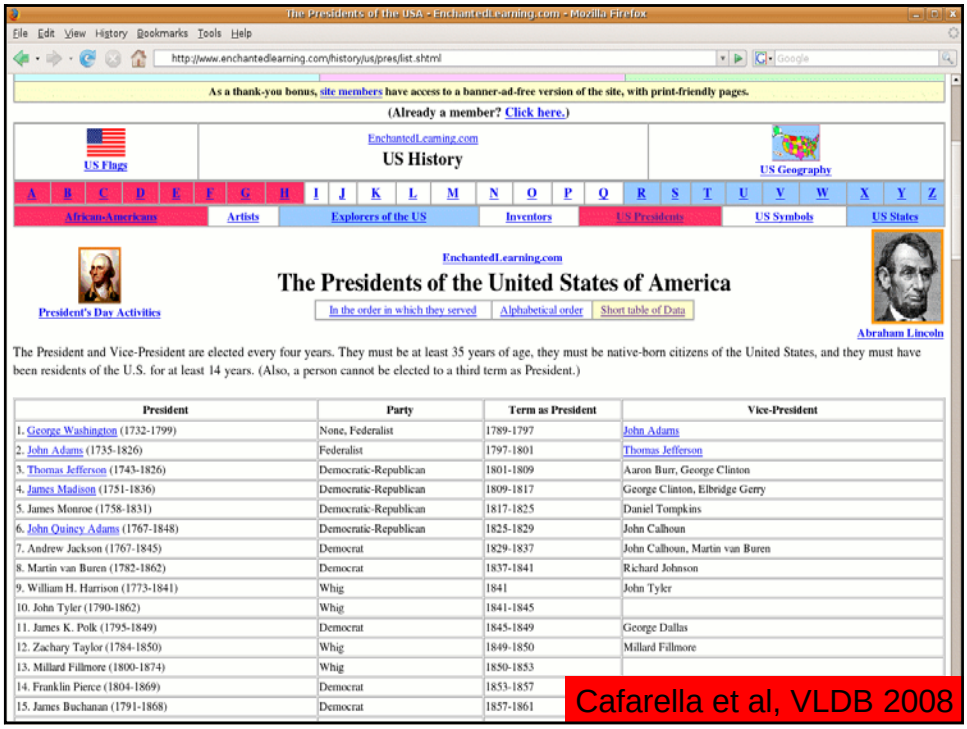
DBLife: a community dataspace

[Doan et al., U. Wisc; CIMPLE, UW + Y!]

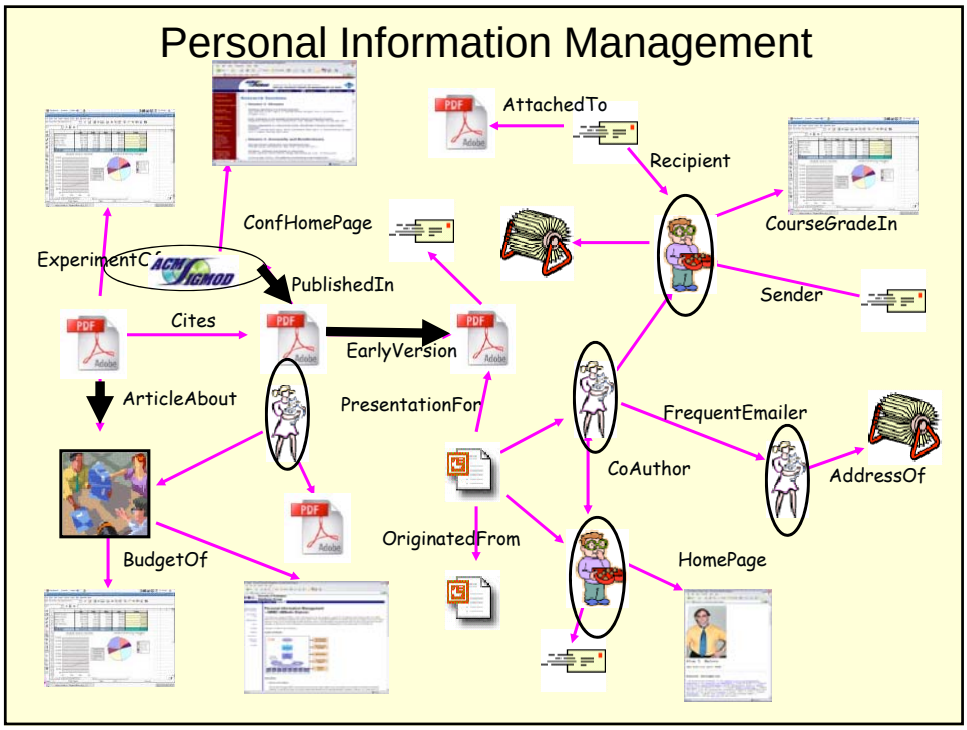
- A data space for the database research community
- Started with 846 data sources in May 2005
  - researcher homepages, CS dept homepages, etc
- Immediately provided some basic service
  - crawl sources daily to obtain 11000+ pages
  - index & provide keyword search
- Incrementally extract & integrate data
  - provide more services & better services
  - leverage user feedback to further improve the system

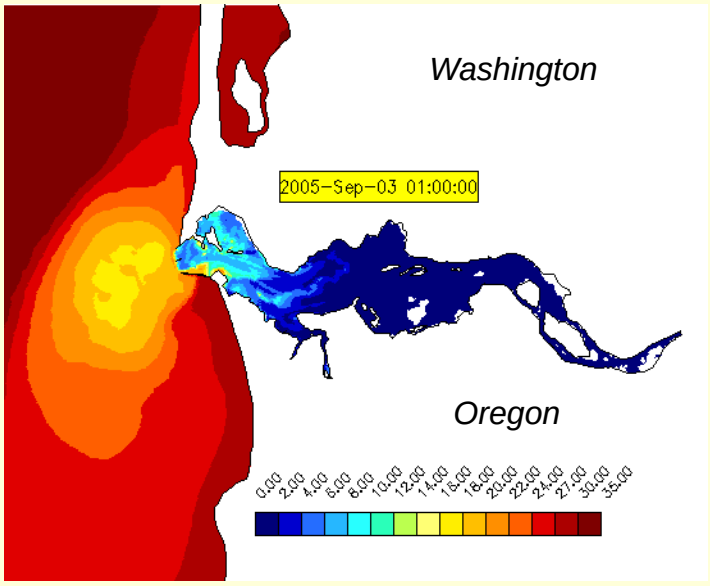
Example Service:

Create SuperHomepages



Cafarella et al, VLDB 2008





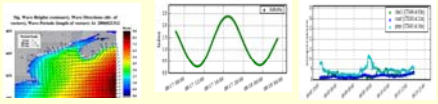
salinity

Example: Ocean Circulation Forecasting System

Atmospheric models

Tides

River Discharge



filesyste

perl and cron

forcings (i.e., inputs)

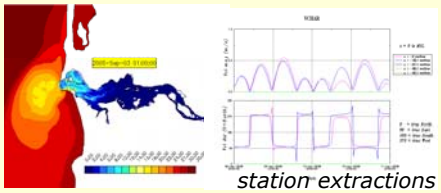
products via the web

FORTRAN

RDBMS

perl and cron

cluster



salinity isolines

station extractions

model-data comparisons

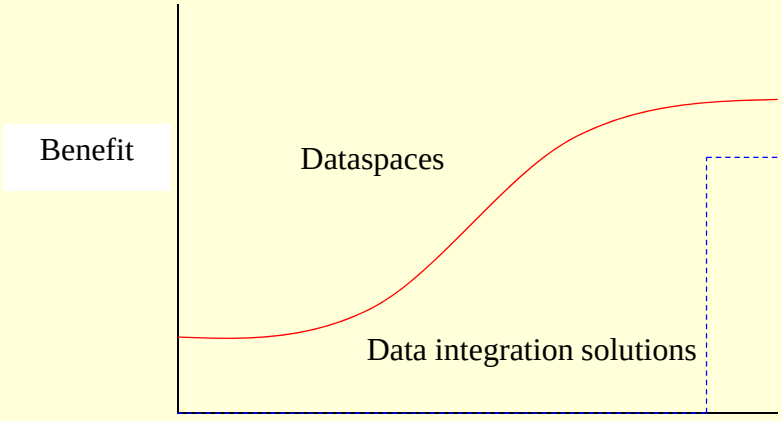
Simulation results
Config and log files
Intermediate files
Annotations
Data Products
Relations
Code

Courtesy Bill Howe, Center for Coastal Margin Observation and Prediction

System Desiderata

- Manage all the data I have, not only what I explicitly put into it via a defined schema
  - Start working with data *where is, as is*.
- Pay-as-you-go:
  - Little or no setup time.
  - Provide *best-effort* services
  - Infers semantics to improve services. Gets help where it's most useful.

Key Principle



Artist: Mike Franklin

Dataspaces:
A System and a Philosophy

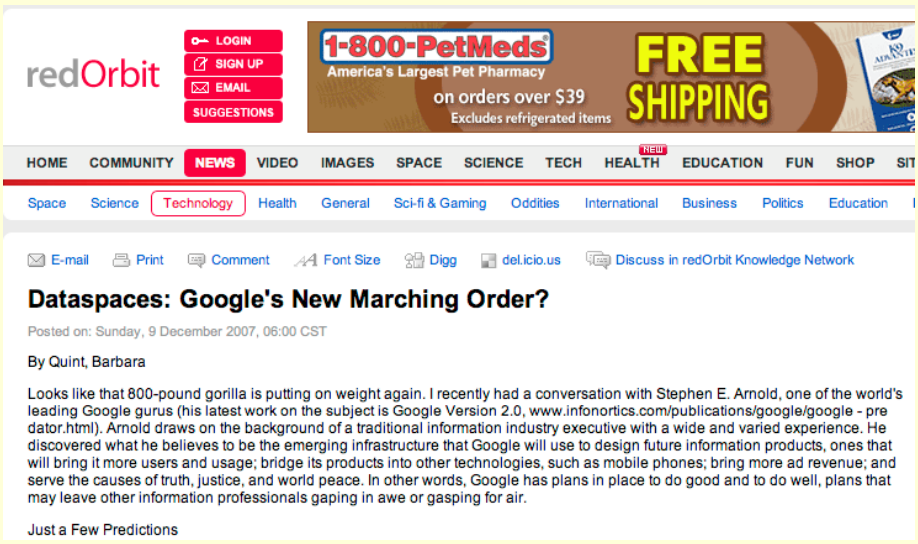
- New kinds of systems:
  - A new kind of data integration system
  - Online data sharing and discussion systems
- Principles applicable anywhere:
  - Best-effort querying, consistency, …
  - Pay-as-you-go data management
- This tutorial: introduces the principles in the context of data integration.

The Buzz on the Street

A significant long-term goal for our community is to transition from managing traditional databases consisting of well-defined schemata for structured business data, to the much more challenging task of managing a rich collection of structured, semi-structured and unstructured data, spread over many repositories in the enterprise and on the Web. This has sometimes been referred to as the challenge of managing dataspaces.

Surajit Chaudhuri

The Buzz in the Press



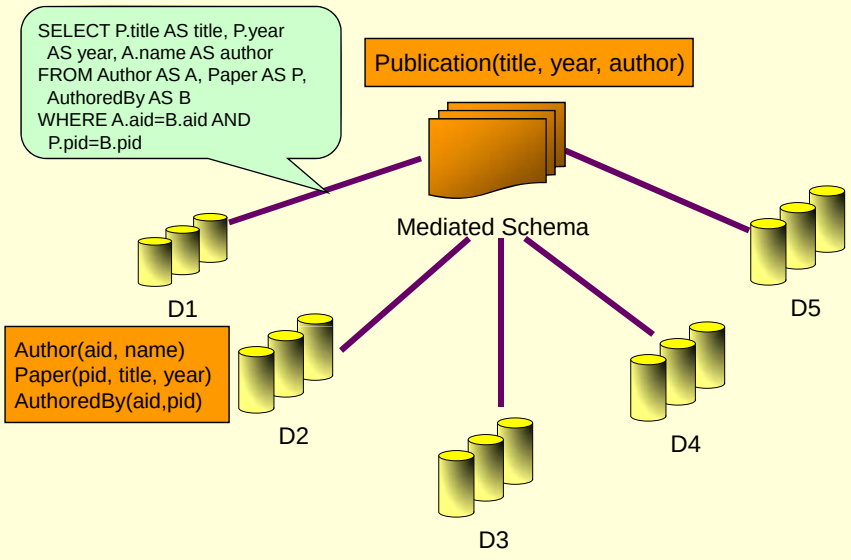
Outline

- ✓ Introduction
- Dataspace principles through data integration:
  - Recap of data integration
  - Locating and understanding data sources
  - Creating mappings and mediated schemas
  - Pay-as-you-go: improving with time
  - Query processing for dataspaces
- Research challenges on specific dataspaces:
  - The Web, Science, the desktop

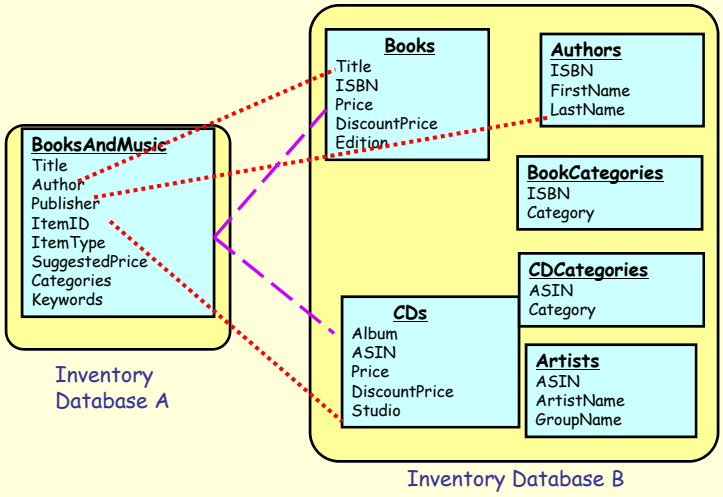
Recap of Data Integration

- Locating, understanding data sources
- Creating a mediated schema
- Creating schema mappings
- Querying and query processing

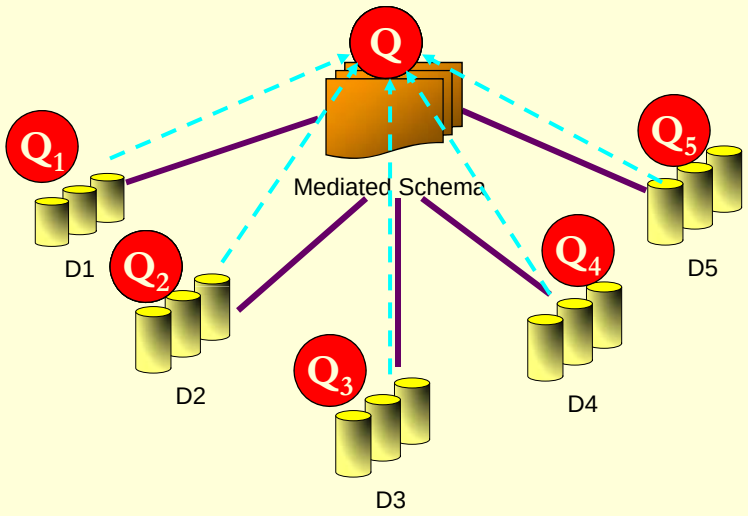
Traditional Data Integration Systems



Semantic Mappings



Querying on Traditional Data Integration Systems



Differences in Dataspaces

[to be covered next]

- Locating and understanding data sources is a challenge
- Semantic mappings will be created automatically:
  - May be approximate
- Create mediated schemas automatically
- Improve semantic mappings over time
- Different query mechanisms
  - Approximate and partial answers
  - User feedback at query stage

The Real Point

- Business data management: done.
- Now: data management = enabling collaboration.
- Collaboration = sharing + combining data, typically in ad-hoc manner.
- Ad-hoc sharing and combining = dataspaces

Finding and Figuring Out Data Sources

Several activities here:

- Identification
- Familiarization
- Characterization
- Customization
- Selection

Steps are interwoven

Source Identification

- Deep-Web Search (e.g., CompletePlanet)
- Brute force?
  - Start with everything you can find (don't worry about relevance)
  - Postpone decisions on inclusion, e.g., until query time
- No generic, automated solution yet
  - How do you find relevant data that is "laying around", e.g., in a spreadsheet on someone's laptop?

Other People's Data

Why is other people's data hard to understand?

- They aren't around to explain it
- What they wrote down about it wasn't quite true or complete in the first place
- The kinds of data in a source have expanded beyond the original intent: *schema drift*

How do you understand the data?

25

Example Dataspace: RxSafe

Consolidated medication list for rural elders

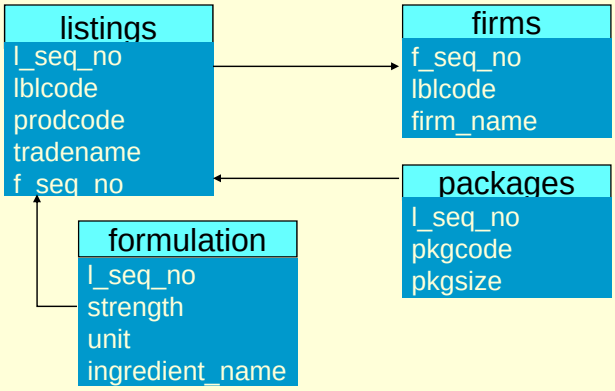
Points in lifetime of a prescription

- Order (clinic, hospital)
- Dispensing (pharmacy)
- Approval (insurer)
- Administration (rehabilitation facility)

Relevant Standards

NDCD, RxNorm, NDF-RT

NDCD:
National Drug Code Directory



Codes for drug packages

Sample NDC: 62584-023-00

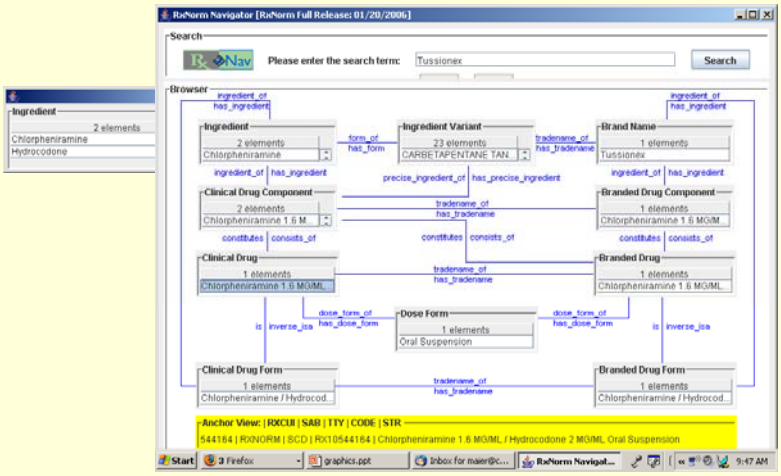
l_seq_no	lblcode	prodcode	tradename	f_seq_no
172062	***62584***	***023***	Vicodin tab	59064

l_seq_no	strength	unit	ingredient_name
172062	5	MG	Hydrocodone
172062	500	MG	Acetaminophen

l_seq_no	pkgcode	pkgsize
172062	***00***	100

f_seq_no	lblcode	firmname
59064	62584	Amerisource

RxNorm: Drug Nomenclature



RxNav from National Library of Medicine

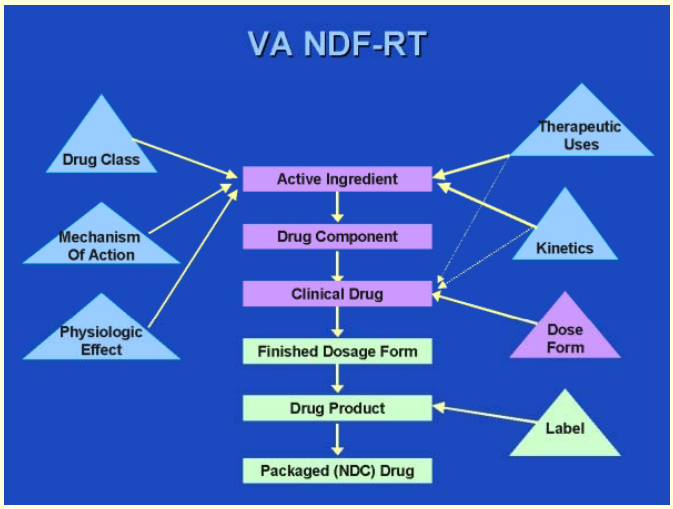
NDF-RT

National Drug File – Reference Terminology

From Veterans Affairs

- Drug class
- Chemical class
- Effects and actions

NDF-RT (Light Blue)



Goals of Understanding Based on Intended Purpose

- Grouping similar medications
- Connecting possible incarnations of same prescription
  - Generic – Brand Name
- Combining medication information for a given patient
  - Must be ***error preserving***

What You Know is Wrong

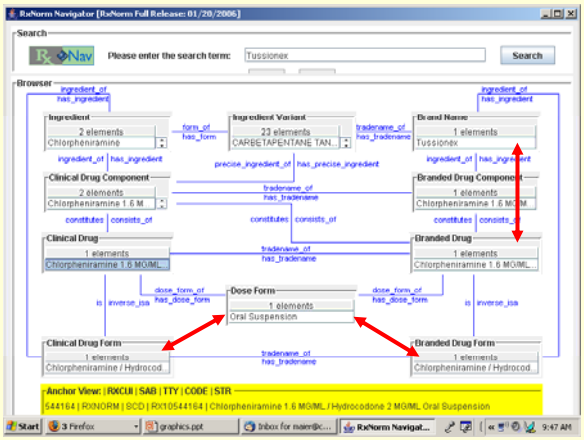
NDC and RxNorm talking about same things

- NDC tradenames: 18913
- RxNorm brand names: 7600
- Strings in common: 418

All RxNorm relationships have explicit inverses

33

What You Know is Incomplete



Doesn't mention atoms, attributes

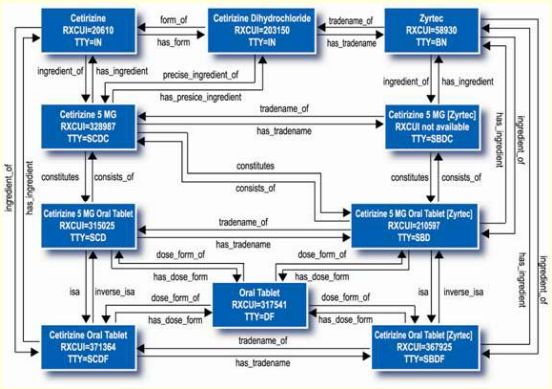
Doesn't include SY, ET, OCD, OBD

Source: National Library of Medicine

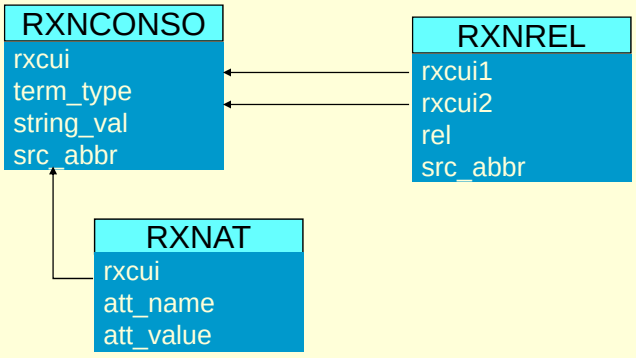
You Might Misunderstand What You're Told

RxNorm diagram is for instances

Multi-ingredient drug case not covered



The Structure Isn't What You Thought



RxNorm uses UMLS, not domain-specific

More complex than this – can have several *atoms* in each concept

How Do You Make Progress?

- Tools with minimal assumptions
  that can help familiarize yourself with the data
- Ability to check hypothesis
  to characterize what's true about the data (or almost true)
- Means to customize data to intended task
  once you understand something about it

What I Want: Dataspace Charting Toolkit

Help with *Familiarization, Profiling, Enhancement*

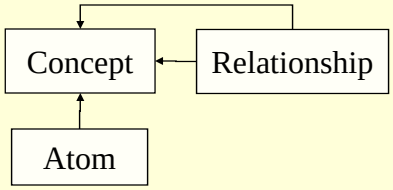
- Inspector for generic models
- Dataspace profiler
  - Assumption tracker and checker
  - Structure discovery techniques
- Customization to task based on discovered characteristics

Quarry: Scalable, Schema-less Browsing

Data Model [Howe, Rayner+ IIMAS 08]

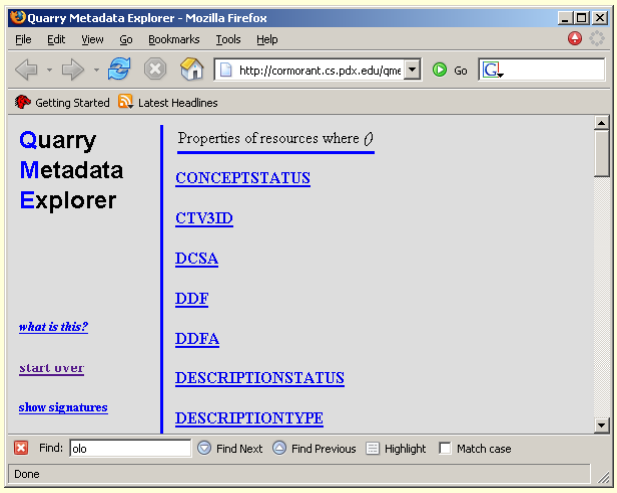
- resource, property, value
  (subject, predicate, object) if you prefer
- no intrinsic distinction between literal values and resource values
  can't necessarily tell the difference initially
- no explicit types or classes

Example: RxNorm

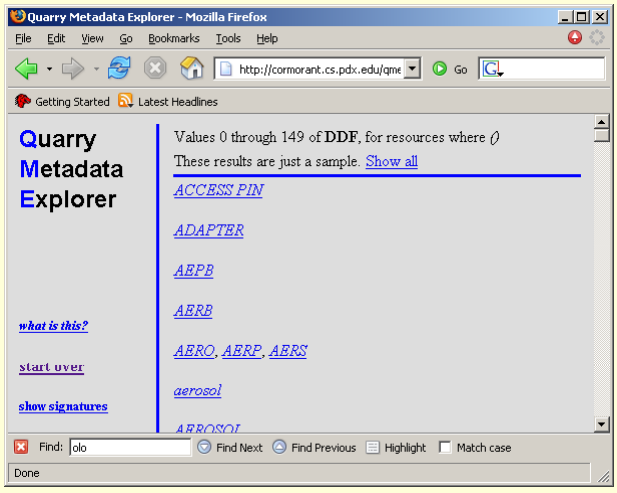


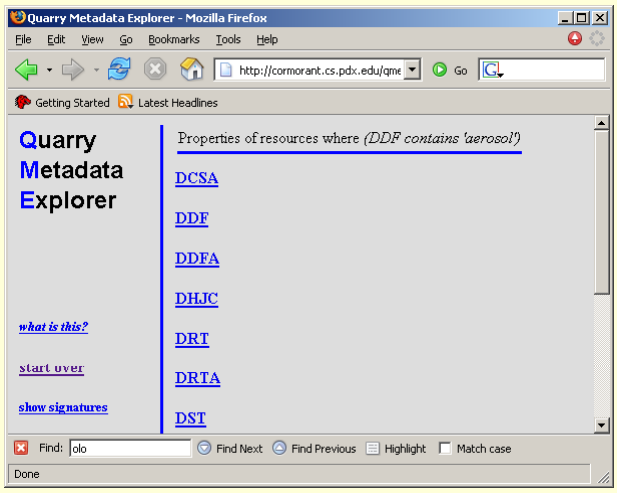
userkey	prop	value
10001	NDC	1
10001	ORIG_CODE	123
10001	ingredient_of	10004
10001	type	DC

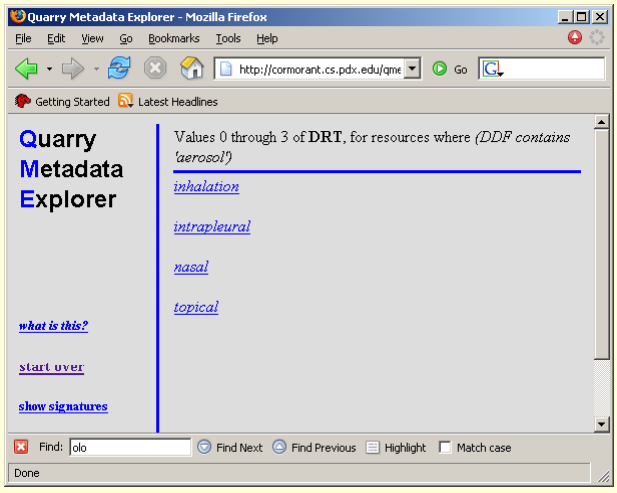
up to 23M triples describing 0.6M concepts and atoms

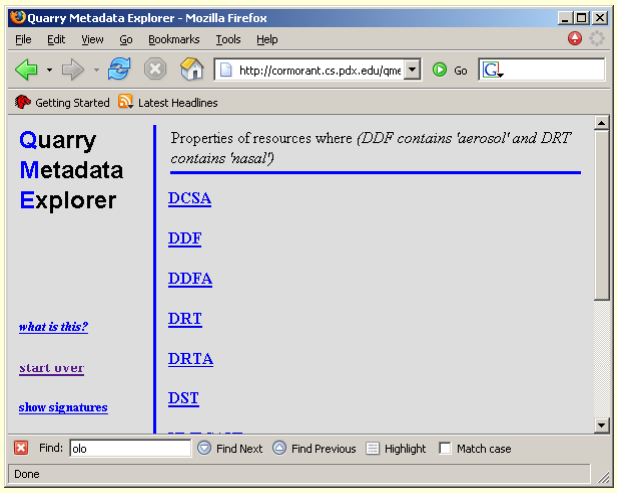

Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
Quarry Metadata Explorer
Properties of resources where ()
CONCEPTSTATUS
CTV3ID
DCSA
DDF
DDFA
DESCRIPTIONSTATUS
DESCRIPTIONTYPE
what is this?
start over
show signatures
Find: olo
Find Next
Find Previous
Highlight
Match case
Done

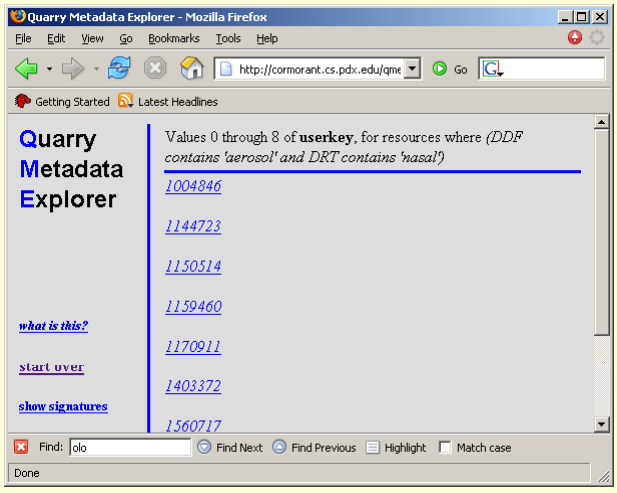
SKIP

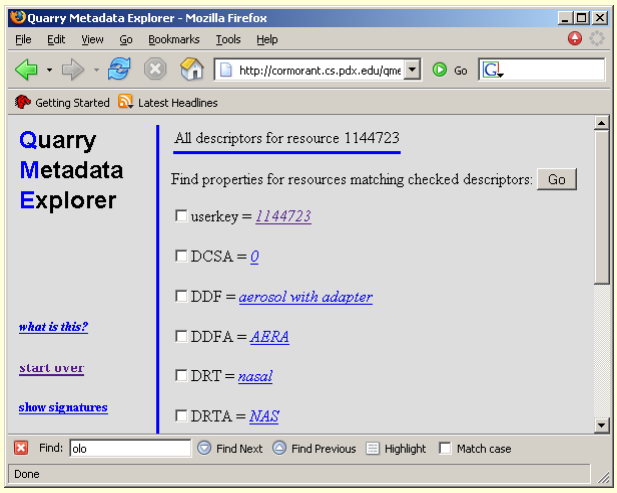

Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
Quarry Metadata Explorer
Values 0 through 149 of DDF, for resources where ()
These results are just a sample. Show all
ACCESS PIN
ADAPTER
AEPB
AERB
AERO, AERP, AERS
aerosol
what is this?
start over
show signatures
Find: olo
Find Next
Find Previous
Highlight
Match case
Done

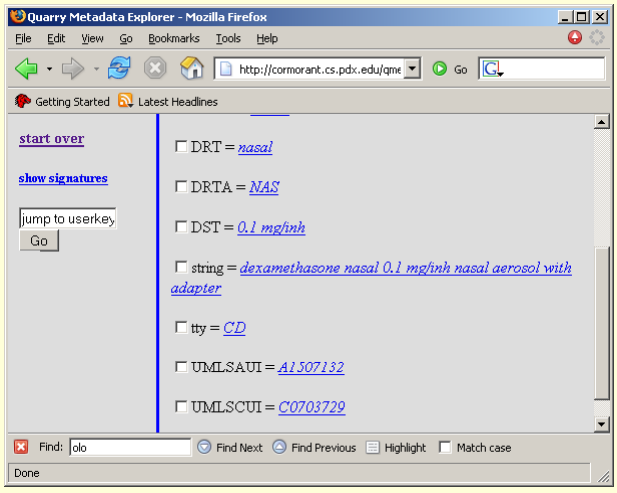

Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
Quarry Metadata Explorer
Properties of resources where (DDF contains 'aerosol')
DCSA
DDF
DDFA
DHJC
DRT
DRTA
DST
what is this?
start over
show signatures
Find: olo
Find Next
Find Previous
Highlight
Match case
Done


Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
Quarry Metadata Explorer
Values 0 through 3 of DRT, for resources where (DDF contains 'aerosol')
inhalation
intrapleural
nasal
topical
what is this?
start over
show signatures
Find: olo
Find Next
Find Previous
Highlight
Match case
Done

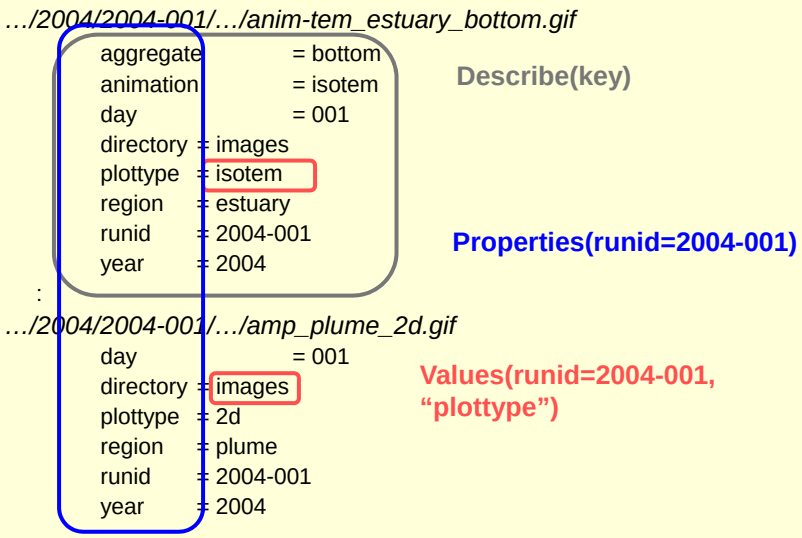

Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
Quarry Metadata Explorer
Properties of resources where (DDF contains 'aerosol' and DRT contains 'nasal')
DCSA
DDF
DDFA
DRT
DRTA
DST
what is this?
start over
show signatures
Find: olo
Find Next
Find Previous
Highlight
Match case
Done


Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
Quarry Metadata Explorer
Values 0 through 8 of userkey, for resources where (DDF contains 'aerosol' and DRT contains 'nasal')
1004846
1144723
1150514
1159460
1170911
1403372
1560717
what is this?
start over
show signatures
Find: olo
Find Next
Find Previous
Highlight
Match case
Done


Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
Quarry Metadata Explorer
what is this?
start over
show signatures
All descriptors for resource 1144723
Find properties for resources matching checked descriptors:
Go
userkey = 1144723
DCSA = 0
DDF = aerosol with adapter
DDFA = AERA
DRT = nasal
DRTA = NAS
Find: olo
Find Next
Find Previous
Highlight
Match case
Done


Quarry Metadata Explorer - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://cormorant.cs.pdx.edu/qme
Go
Getting Started
Latest Headlines
start over
show signatures
jump to userkey
Go
DRT = nasal
DRTA = NAS
DST = 0.1 mg/inh
string = dexamethasone nasal 0.1 mg/inh nasal aerosol with adapter
tty = CD
UMLSAUI = A1507132
UMLSCUI = C0703729
Find: olo
Find Next
Find Previous
Highlight
Match case
Done

Quarry API



Behind the Scenes

Signatures

- resources possessing the same properties clustered together
- Posit that |Signatures| << |Resources|
- Queries evaluated over Signature Extents

Dataspace Profiling

Commercial profilers: DataFlux, Infogix (ACR), KnowledgeDriver
- prep for cleaning, migration
- generally relational model, by table

Potter's Wheel [Raman, Hellerstein VLDB 2001]
- learning column transformation
- unfolding – data value to column labels

Cross-Source Profiles

Bellman [Dasu, Johnson+ SIGMOD 02]
- Find joinable columns (1-N, M-N)
- Is one field the composition of others?
- Part of T joins with T1, part with T2

Make the point that database schemas "devolve" with time as business processes change

52

Dataspace Profiling: NDC Examples

- l_seq_no is key of listings – yes
- lblcode, prodcode key of listings – no

  45,953 45,972 (19)
- firm_name → lblcode – no

  2931 2952 (21)
- each product listing should have >0 packages and >0 ingredients

  44,972 (1180) 45,180 (792)

Courtesy of Nick Rayner

Checking Across Sources NDC vs. RxNorm

Ingredients

- 2794 ingredient names in NDCD
- 5145 ingredients in RxNorm
- 1570 equal strings

What to Do with Flawed Assumptions?

- Track exceptions
- Refine assumption

  firm_name, location → lblcode
- Refine knowledge of world

  RxNorm has *ingredient variants* (which have the same type as ingredients)
- Want to track assumptions as they evolve, results of checks

Customization: InfoSonde

Support customizations appropriate to discovered data characteristics
[Howe, Rayner+ IIMAS 2008]

Three-part modules

- Probe: Check or discover properties
- Switch: Present applicable customizations
- Check: Test that chosen switch is still valid

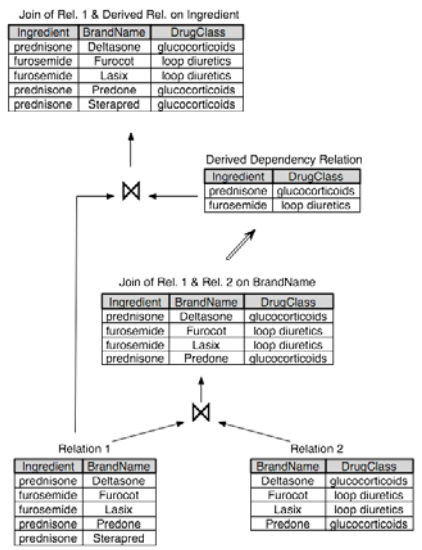
Functional Dependency Module

- Probe: Test for FD
- Switch:
  - FD holds: add constraint, decompose
  - FD fails: partition, repair
- Check:

  Example – if using decompose, check that FD still holds

57

Linkage Extension Module



Want to extend a linkage based on a discovered functional relationship

Probe: Join satisfies FD
here, ING → DC

Switch: Materialize functional relationship
can be used to extend original relation with DC

Check: Test that FD still holds

Related Work: Characterization and Enhancement

Structure Discovery [Andritsos, Miller+ SIGMOD 2005]

- Clustering of columns, rows
- Ranking FDs by corresponding redundancy

59

Other Structural Work

- Wide-table workbench [Chu, Baid+ VLDB 07]
  - Start with unstructured records & keyword search
  - Extract, Integrate, Cluster to get more structure
- LearnPADs [Fisher, Walker+ POPL 08]
  - Targets *ad hoc data*: semi-structured data with no toolset (e.g., log file)
  - Induces PADS grammar from examples, use to instantiate generic tools (parser, accumulator, formatter, translator)

Source Selection

- Obviously, familiarization and characterization can help in evaluating a single source
- But you may want to judge a combination of sources

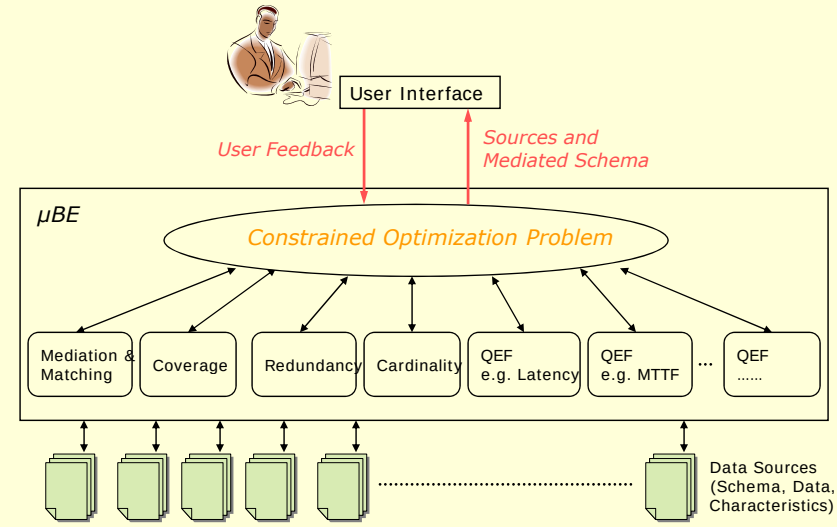
µBE: User Guided Source Selection

Matching by Example

[Aboulnaga, El Gebaly ICDE 2007]

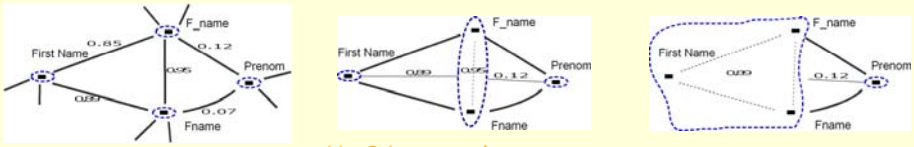
- Want to select set of sources based on matching quality, cardinality, coverage, redundancy and possibly other measures (latency)
- Mediated schema consists of global attributes = sets of source attributes
- User can constrain sources and global attributes

µBE Architecture

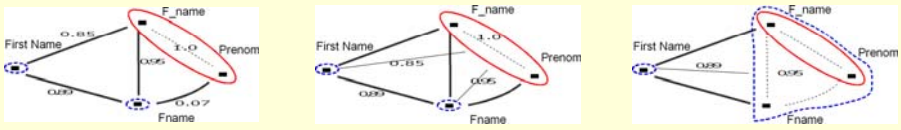


Courtesy Ashraf Aboulnaga & Kareem El Gebaly

µBE Matching Example



No GA constraints



With a GA constraint

Courtesy Ashraf Aboulnaga & Kareem El Gebaly