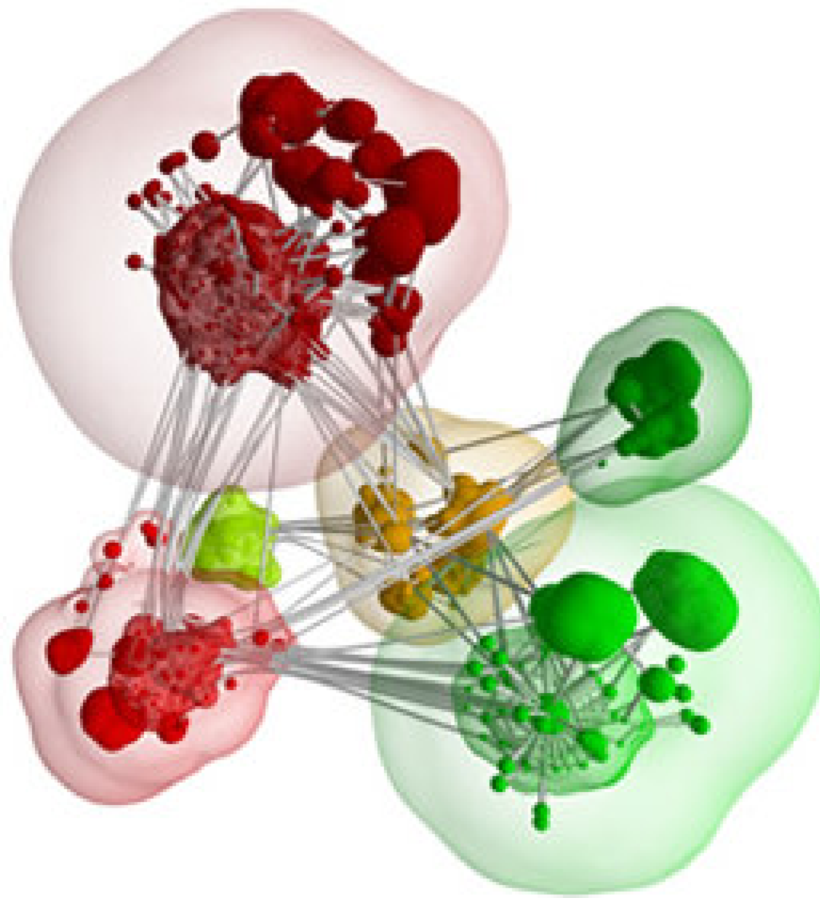
# EDW2010
# Data Integration with Semantic Web Tools



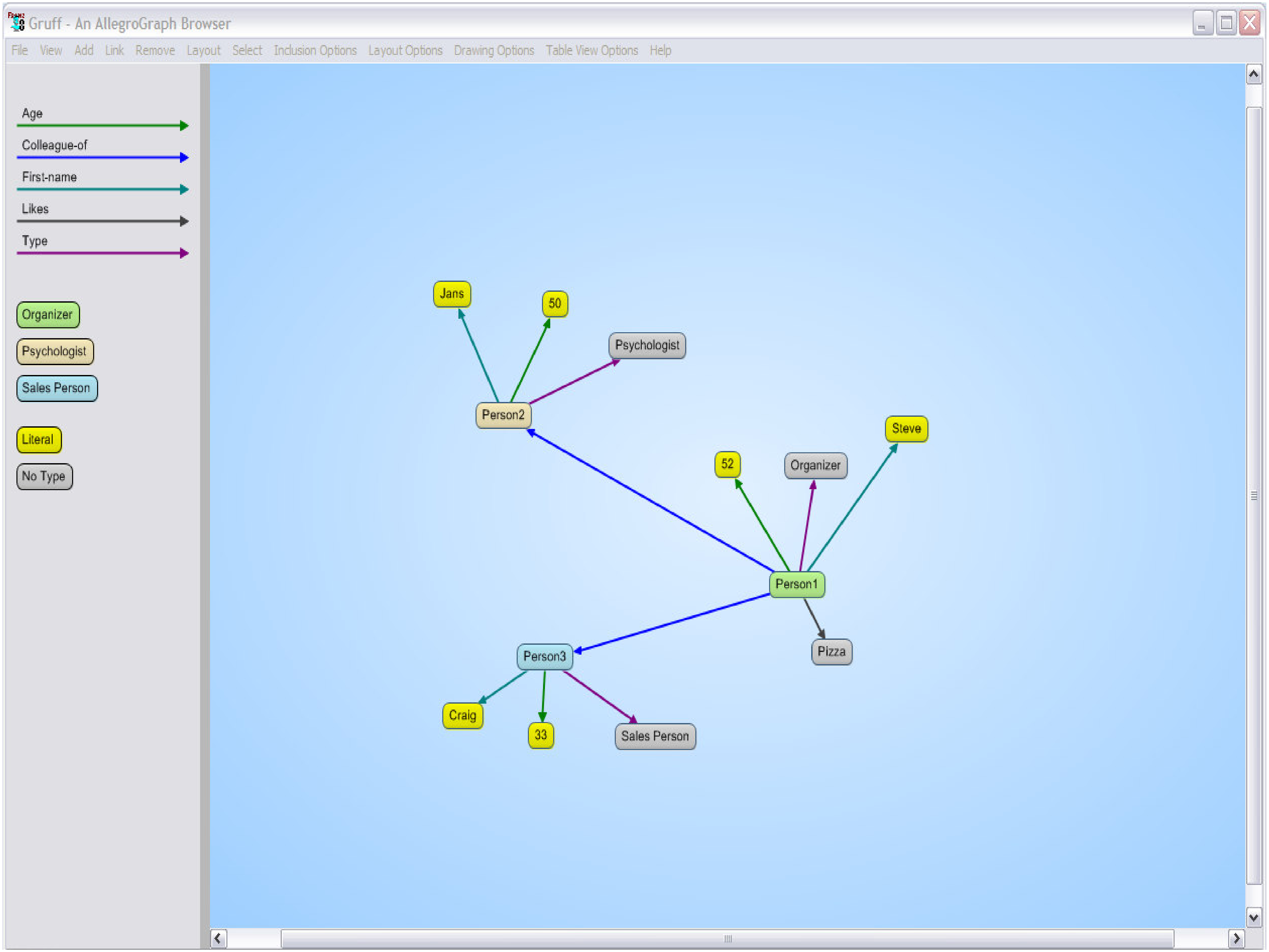
Jans Aasman, Ph.D.
CEO Franz Inc
Ja@Franz.com

# Contents

- We integrate through a set of tools in our triple store
- A three minute introduction to triple stores
- Data integration with Linked Open Data [Demo]
- Can we do this integration in the RDB world?
  - Pfizer, BC
- Our current process for organic data integration
  - Vocabularies, Thesauruses, Taxonomy, Ontologies
  - Schema Spaces
  - RDFy-ing your data (kind of ETL)
  - Matching your data and building an inverted metadata instance store
  - Querying

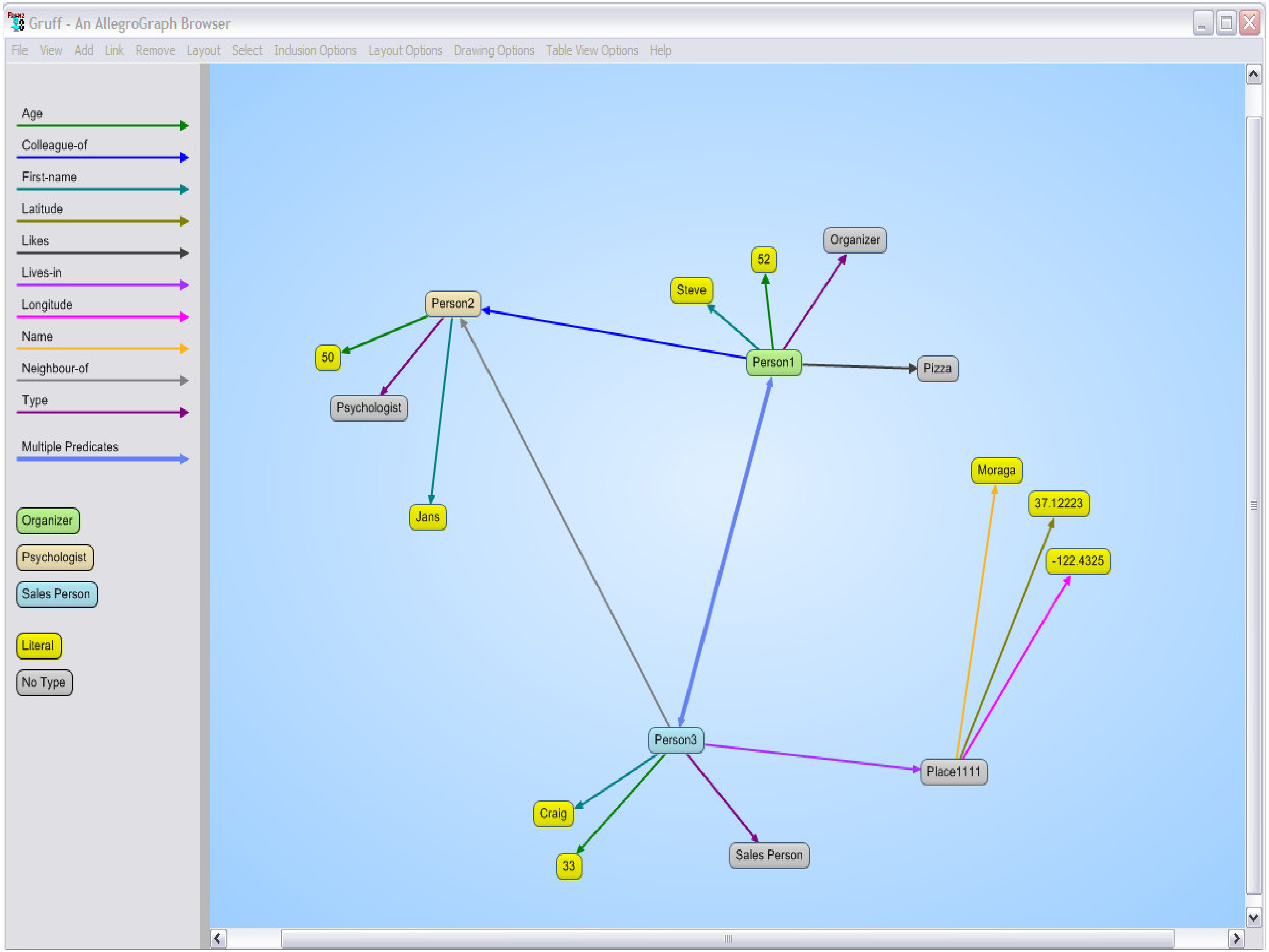
## Graphs, triples, triple-store?

```
createTripleStore("seminar.db" )

addTriple (Person1 first-name Steve)
addTriple (Person1 isa  Organizer)
addTriple (Person1 age 52)
addTriple (Person2 first-name Jans)
addTriple (Person2 isa Psychologist)
addTriple (Person2 age 50)
addTriple (Person3 first-name Craig)
addTriple (Person3 isa SalesPerson)
addTriple (Person3 age 32)

addTriple (Person1 colleague-of Person2)
addTriple (Person1 colleague-of Person3)

addTriple (Person1 likes Pizza)
```


Gruff - An AllegroGraph Browser
File View Add Link Remove Layout Select Inclusion Options Layout Options Drawing Options Table View Options Help
Age
Colleague-of
First-name
Likes
Type
Organizer
Psychologist
Sales Person
Literal
No Type
Jans
50
Psychologist
Person2
Steve
52
Organizer
Person1
Pizza
Person3
Craig
33
Sales Person

# Keep adding New Relationships

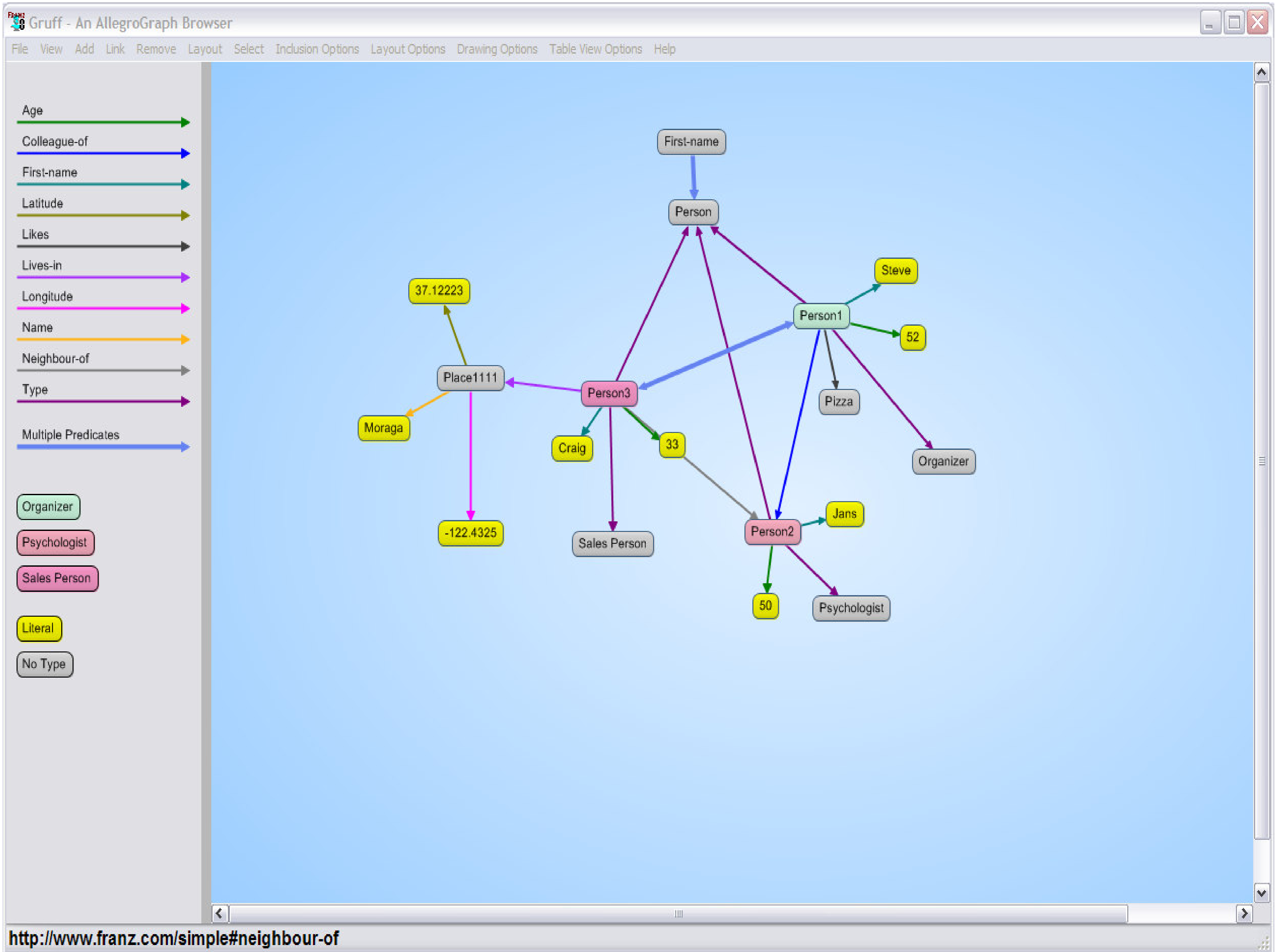
```
addTriple ( Person3 neighbor-of Person1)
addTriple ( Person3 neighbor-of Person2)

addTriple ( Person3 !o:lives-in !o:Place1111)
addTriple ( Place1111 !o:name !"Moraga")
addTriple ( Place1111 !o:latitude !"37.12223")
addTriple ( Place1111 !o:longitude !"-122.4325")
```

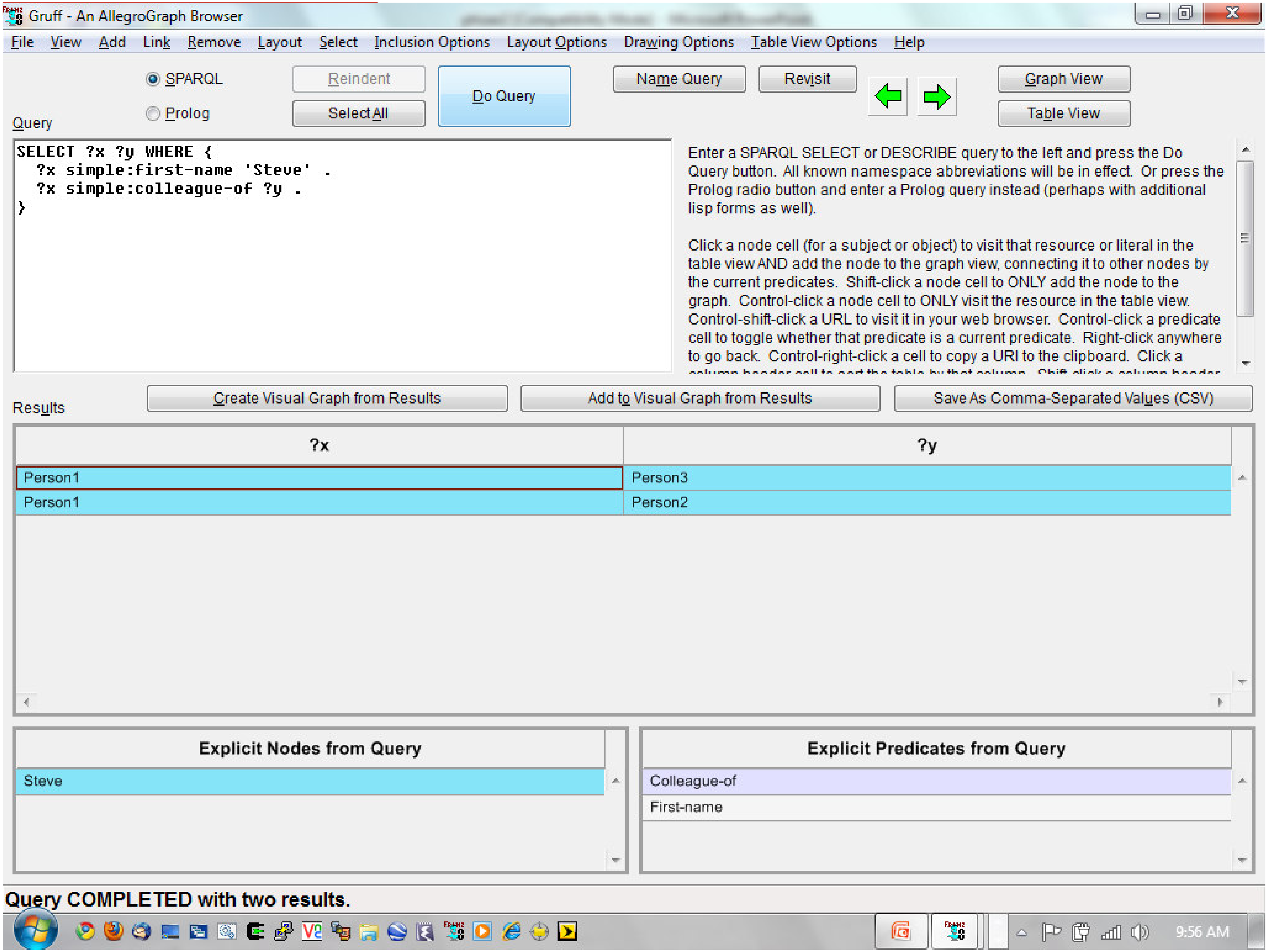

Gruff - An AllegroGraph Browser
File View Add Link Remove Layout Select Inclusion Options Layout Options Drawing Options Table View Options Help
Age
Colleague-of
First-name
Latitude
Likes
Lives-in
Longitude
Name
Neighbour-of
Type
Multiple Predicates
Organizer
Psychologist
Sales Person
Literal
No Type
Organizer
52
Steve
Person2
50
Person1
Pizza
Psychologist
Jans
Moraga
37.12223
-122.4325
Person3
Place1111
Craig
33
Sales Person

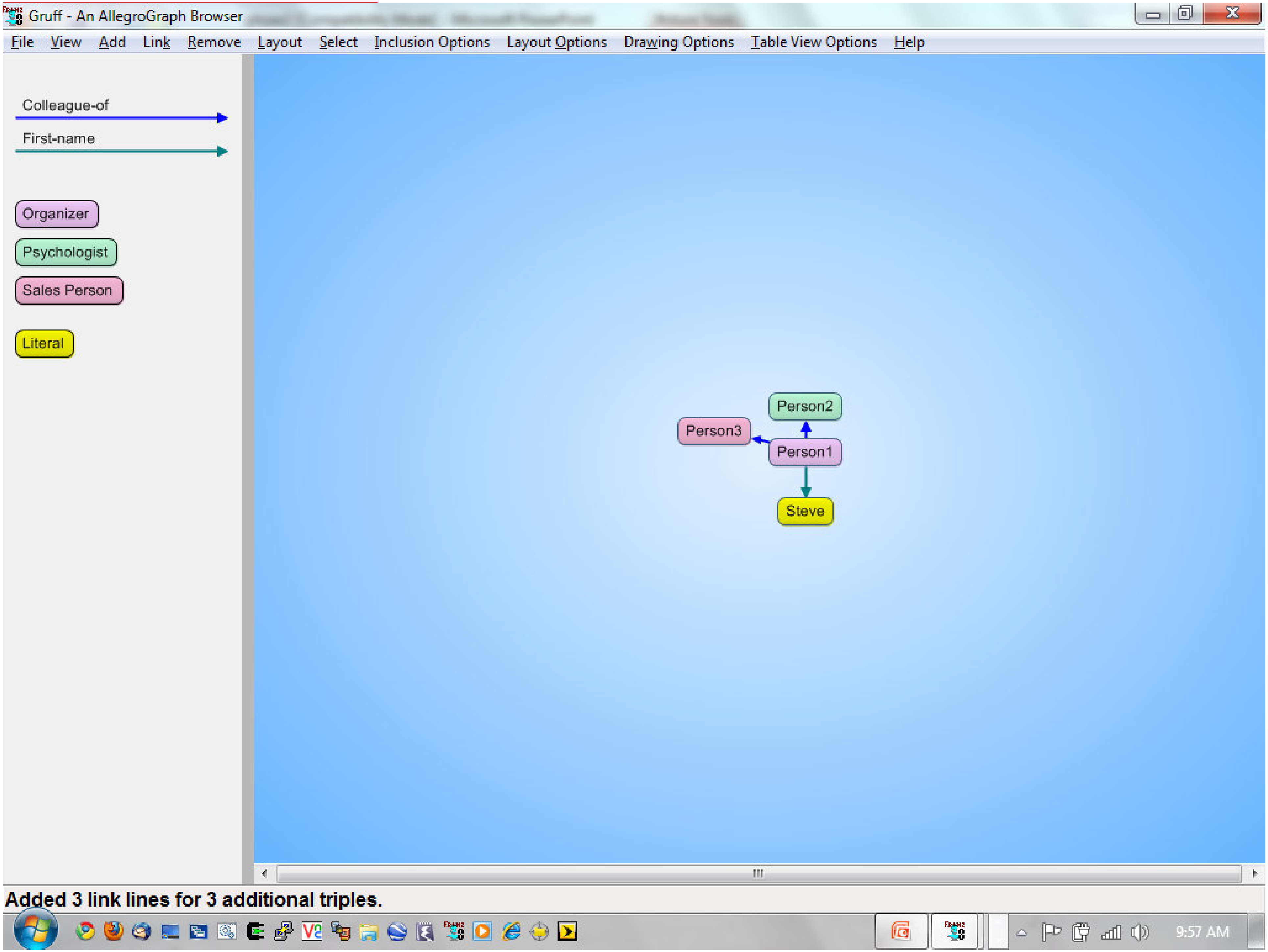
# Apply Logic – Infer New Relationships

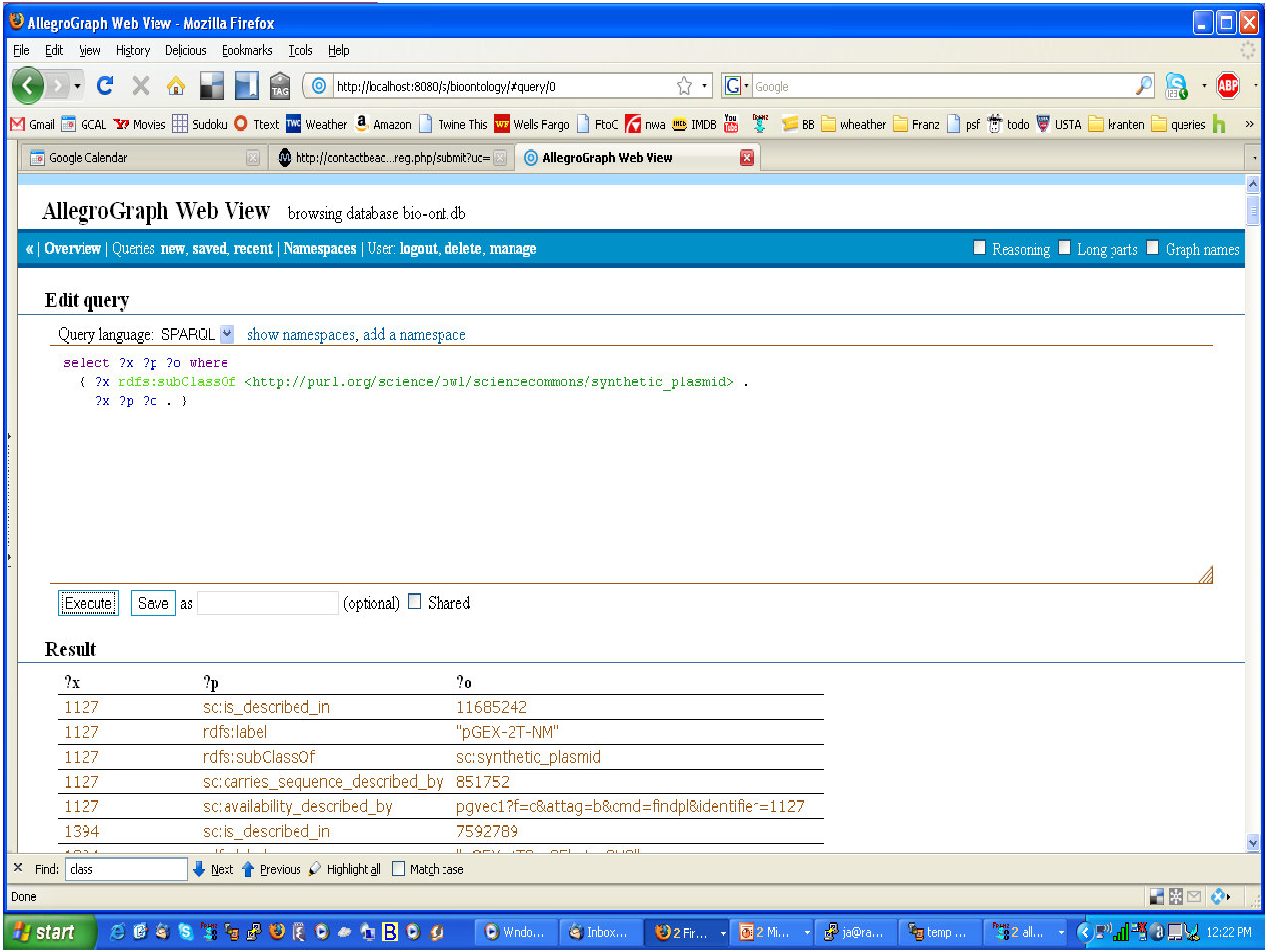
```
addTriple (first-name domain Person)
```

Every thing that has a first name must be a person


Gruff - An AllegroGraph Browser
File View Add Link Remove Layout Select Inclusion Options Layout Options Drawing Options Table View Options Help
Age
Colleague-of
First-name
Latitude
Likes
Lives-in
Longitude
Name
Neighbour-of
Type
Multiple Predicates
Organizer
Psychologist
Sales Person
Literal
No Type
First-name
Person
Steve
Person1
52
37.12223
Place1111
Person3
Pizza
Moraga
Craig
33
Organizer
-122.4325
Sales Person
Person2
Jans
50
Psychologist
http://www.franz.com/simple#neighbour-of


Gruff - An AllegroGraph Browser
File View Add Link Remove Layout Select Inclusion Options Layout Options Drawing Options Table View Options Help
SPARQL
Prolog
Reindent
Select All
Do Query
Name Query
Revisit
Graph View
Table View
Query
SELECT ?x ?y WHERE {
?x simple:first-name 'Steve' .
?x simple:colleague-of ?y .
}
Enter a SPARQL SELECT or DESCRIBE query to the left and press the Do Query button. All known namespace abbreviations will be in effect. Or press the Prolog radio button and enter a Prolog query instead (perhaps with additional lisp forms as well).
Click a node cell (for a subject or object) to visit that resource or literal in the table view AND add the node to the graph view, connecting it to other nodes by the current predicates. Shift-click a node cell to ONLY add the node to the graph. Control-click a node cell to ONLY visit the resource in the table view. Control-shift-click a URL to visit it in your web browser. Control-click a predicate cell to toggle whether that predicate is a current predicate. Right-click anywhere to go back. Control-right-click a cell to copy a URI to the clipboard. Click a
Results
Create Visual Graph from Results
Add to Visual Graph from Results
Save As Comma-Separated Values (CSV)
?x
?y
Person1
Person3
Person1
Person2
Explicit Nodes from Query
Steve
Explicit Predicates from Query
Colleague-of
First-name
Query COMPLETED with two results.
9:56 AM


Gruff - An AllegroGraph Browser
File View Add Link Remove Layout Select Inclusion Options Layout Options Drawing Options Table View Options Help
Colleague-of
First-name
Organizer
Psychologist
Sales Person
Literal
Person2
Person3
Person1
Steve
Added 3 link lines for 3 additional triples.
9:57 AM


AllegroGraph Web View - Mozilla Firefox
File Edit View History Delicious Bookmarks Tools Help
http://localhost:8080/s/bioontology/#query/0
Gmail GCAL Movies Sudoku Ttext Weather Amazon Twine This Wells Fargo FtoC nwa IMDB BB wheather Franz psf todo USTA kranten queries
Google Calendar
http://contactbeac...reg.php/submit?uc=
AllegroGraph Web View
AllegroGraph Web View browsing database bio-ont.db
« | Overview | Queries: new, saved, recent | Namespaces | User: logout, delete, manage
Reasoning Long parts Graph names
Edit query
Query language: SPARQL show namespaces, add a namespace
select ?x ?p ?o where
{ ?x rdfs:subClassOf <http://purl.org/science/owl/sciencecommons/synthetic_plasmid> .
?x ?p ?o . }
Execute Save as (optional) Shared
Result
?x ?p ?o
1127 sc:is_described_in 11685242
1127 rdfs:label "pGEX-2T-NM"
1127 rdfs:subClassOf sc:synthetic_plasmid
1127 sc:carries_sequence_described_by 851752
1127 sc:availability_described_by pgvec1?f=c&attag=b&cmd=findpl&identifier=1127
1394 sc:is_described_in 7592789
Find: class Next Previous Highlight all Match case
Done
start
Windo... Inbox... 2 Fir... 2 Mi... ja@ra... temp ... 2 all...
12:22 PM

W3C

Views: desktop mobile print

W3C By Region Go

STANDARDS PARTICIPATE MEMBERSHIP ABOUT W3C

Google

Skip

## STANDARDS

- Web Design and Applications
- Web Architecture
- Semantic Web
- XML Technology
- Web of Services
- Web of Devices
- Browsers and Authoring Tools
- ... or view all

## WEB FOR ALL

- W3C A to Z
- Accessibility
- Internationalization
- Mobile Web
- eGovernment

### W3C Invites Implementation of Widget Packaging and Configuration

01 December 2009 | Archive

The Web Applications Working Group invites implementation of the Candidate Recommendation of Widget Packaging and Configuration. This specification standardizes a packaging format for software known as widgets. Widgets are client-side applications that are authored using Web standards, but whose content can also be embedded into Web documents. The packaging format acts as a container for files used by a widget. The configuration document is an XML vocabulary that declares metadata and configuration parameters for a widget. The steps for processing a widget package describe the expected behavior and means of error handling for runtimes while processing the packaging format, configuration document, and other relevant files. The group plans to track implementations in an implementation report. Learn more about the Rich Web Client Activity.

### Voice Extensible Markup Language (VoiceXML) 3.0 Draft Published

03 December 2009 | Archive

### Last Call: W3C XML Schema Definition Language (XSD) 1.1

03 December 2009 | Archive

### CSS 2D Transforms, Transitions Modules Updated

01 December 2009 | Archive

### Multimodal Architecture and Interfaces (MMI Architecture) Working Draft Published

01 December 2009 | Archive

### W3C Launches HTML5 Japanese Interest Group

The World Wide Web Consortium (W3C) is an international community that develops standards to ensure the long-term growth of the Web. Join groups, and participate in W3C blogs and other discussion. We welcome your help to fulfill the W3C mission: to lead the Web to its full potential.

## JOBS

W3C is seeking a Chief Executive Officer; learn more about job opportunities.

## W3C BLOG

Default Prefix Declaration
18 November 2009 by Henry S. Thompson

W3C community bridges unicorns and werewolves #tpac09
13 November 2009 by Coralie Mercier

W3C Cheatsheet for developers
5 November 2009 by Dominique Hazaël-Massieux

## VALIDATORS AND OTHER SOFTWARE

# SEMANTIC WEB

On this page → technology topics • news • upcoming events and talks

In addition to the classic "Web of documents" W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The ultimate goal of the Web of data is to enable computers to do more useful work and to develop systems that can support trusted interactions over the network. The term "Semantic Web" refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data. Linked data are empowered by technologies such as RDF, SPARQL, OWL, and SKOS.

## Linked Data

The Semantic Web is a Web of data — of dates and titles and part numbers and chemical properties and any other data one might conceive of. RDF provides the foundation for publishing and linking your data. Various technologies allow you to embed data in documents (RDFa, GRDDL) or expose what you have in SQL databases, or make it available as RDF files.

## Vocabularies

At times it may be important or valuable to organize data. Using OWL (to build vocabularies, or "ontologies") and SKOS (for designing knowledge organization systems) it is possible to enrich data with additional meaning, which allows more people (and more machines) to do more with the data.

## Query

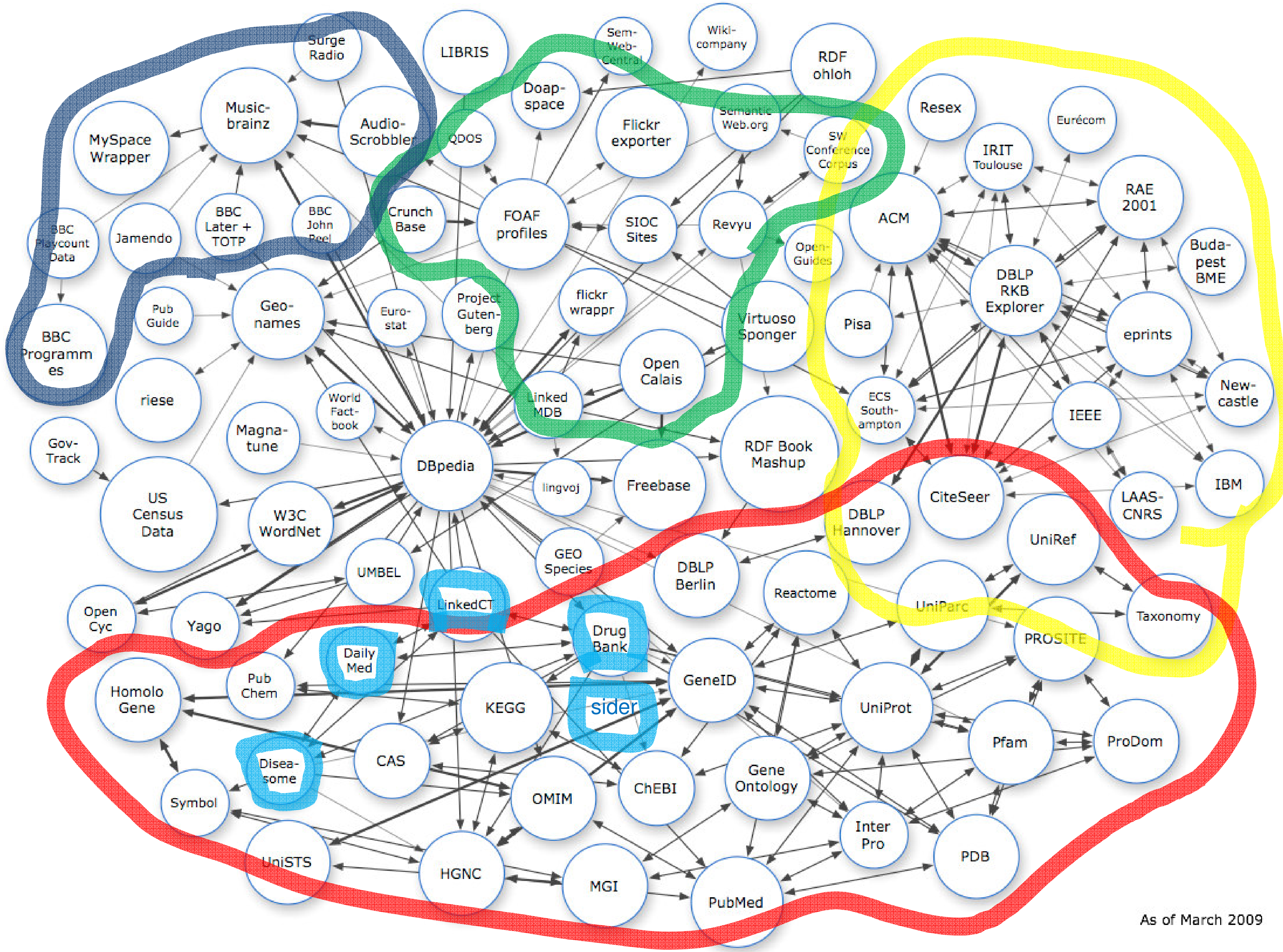
Query languages go hand-in-hand with databases. If the Semantic Web is viewed as a global database, then it is easy to understand why one would need a query language for that data. SPARQL is the query language for the Semantic Web.

## Inference

Near the top of the Semantic Web stack one finds inference — reasoning over data through rules. W3C work on rules, primarily through RIF and OWL, is focused on translating between rule languages and

## Vertical Applications

W3C is working with different industries — for example in Health Care and Life Sciences, eGovernment, and Energy — to improve collaboration, research and development, and innovation adoption



Surge Radio
LIBRIS
Sem-Web-Central
Wiki-company
RDF ohloh
Music-brainz
Audio-Scrobbler
Doap-space
Flickr exporter
Semantic Web.org
Resex
Eurécom
MySpace Wrapper
QDOS
SW Conference Corpus
IRIT Toulouse
RAE 2001
BBC Playcount Data
Jamendo
BBC Later + TOTP
BBC John Peel
Crunch Base
FOAF profiles
SIOC Sites
Revyu
ACM
Open-Guides
Buda-pest BME
BBC Programmes
Pub Guide
Geo-names
Euro-stat
Project Guten-berg
flickr wrappr
Virtuoso Sponger
Pisa
DBLP RKB Explorer
eprints
riese
World Fact-book
Open Calais
Linked MDB
ECS South-ampton
IEEE
New-castle
Gov-Track
Magna-tune
DBpedia
lingvoj
Freebase
RDF Book Mashup
CiteSeer
IBM
US Census Data
W3C WordNet
DBLP Hannover
UniRef
LAAS-CNRS
UMBEL
GEO Species
DBLP Berlin
Reactome
Open Cyc
Yago
LinkedCT
Drug Bank
UniParc
Taxonomy
PROSITE
Daily Med
Homolo Gene
Pub Chem
KEGG
GeneID
UniProt
sider
CAS
Pfam
ProDom
Disea-some
Symbol
OMIM
ChEBI
Gene Ontology
Inter Pro
UniSTS
HGNC
MGI
PubMed
PDB
As of March 2009

# Demoing Data Integration over a federation of 11 linked data sets

- We took 5 public databases: Drugbank, Dailymeds, Clinical trials, Diseasome, and Sider. Entities are mostly linked together through same-as relationships.

- And using some entity extraction created some more databases
  - CT-discusses-drug,
  - CT-discusses-side-effect
  - CT-discusses-target,
  - CT-discusses-disease

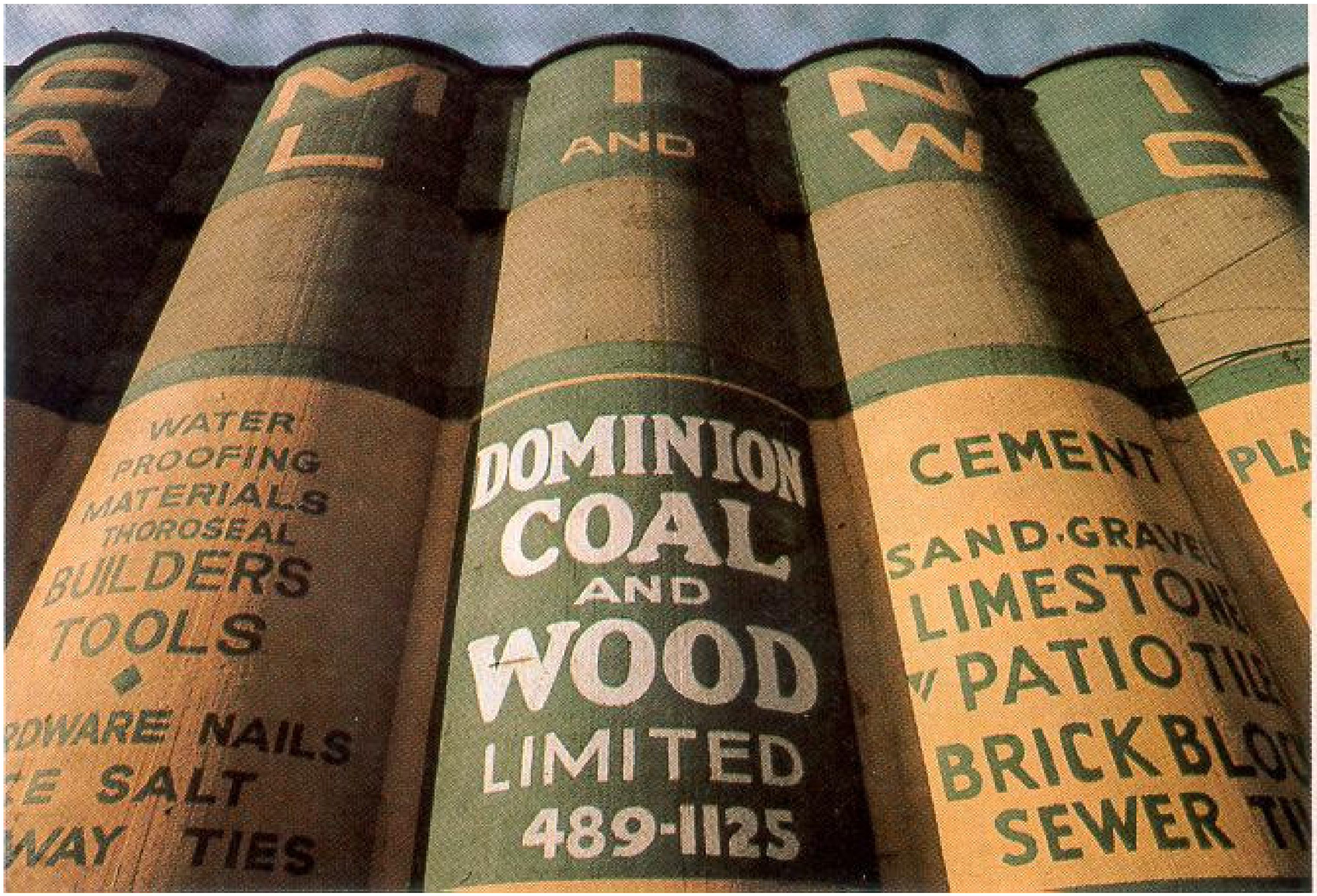
- With some help from Alitora entity extraction on Rheumatoid Arthritis
  - CT-mentions-genes

- And to facilitate search through schema space: Schema-connections

# Interesting queries

- Sparql
  - Give me the title of all clinical trials that discuss the drug Lipitor and the side-effect “Diabetes type 2”
  - Give me clinical trials that discuss Rheumatoid Arthritis and give me the genes and drugs discussed
- Prolog
  - Find all clinical trials that resemble clinical trial NCT00130091 given diseases, drugs, targets, and side-effects

# Can we do this kind of integration in the Relational Database World?



Franz Inc.
"Web 3.0's Database"

Cisco

Raytheon

University of Yamanashi

Mayo Clinic

Adobe

NASA

KT

Tohoku University

gsk GlaxoSmithKline

citi

Lockheed Martin

Noah诺亚舟

Novartis

U.S. Army

Teleonto
Next Generation Revenue Assurance & Analytics

Kodak

KDDI

Johns Hopkins University

GE Global Research
United States - India - China - Germany

Modus Operandi

Lilly

Rensselaer

MITRE

The Redlands Institute

Ford

BAE Systems

National Archives of Korea

Boeing

Duke University

KnowledgeReef

Office of the Director of National Intelligence
United States of America

The University of Texas
Medical School at Houston
A part of The University of Texas Health Science Center at Houston

DISA Defense Information Systems Agency
Department of Defense

Sun microsystems

# Knowledge Sharing using Semantic Technologies

February 25, 2010

Vijay K. Bulusu
*Sr. Manager*
*R&D Informatics*

# Knowledge Sharing via imprecise connections

- **Goal**
  - Identify and aggregate data from various sources in the absence of unique identifiers and lack of referential integrity

- **Challenges**
  - Incompatible databases
    - Same name; different meaning (Batch / Lot Information)
  - Imprecise Connections
    - Lack of controlled vocabulary for key fields
    - One identifier mapped to multiple entities

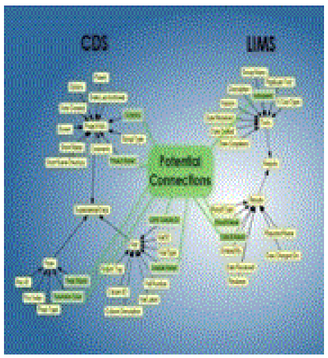
# Business Problem #1

- **Compound Purity Verification**
  - For release results in LIMS for a clinical batch with x% specified/unspecified impurity, FDA wants confirmation on the integration of the peak in the CDS and the calculations of standards and samples to get the final result
  - For an impurity value recorded in LIMS (Certificate of Analysis or Stability Report), find the corresponding impurity value in Empower.

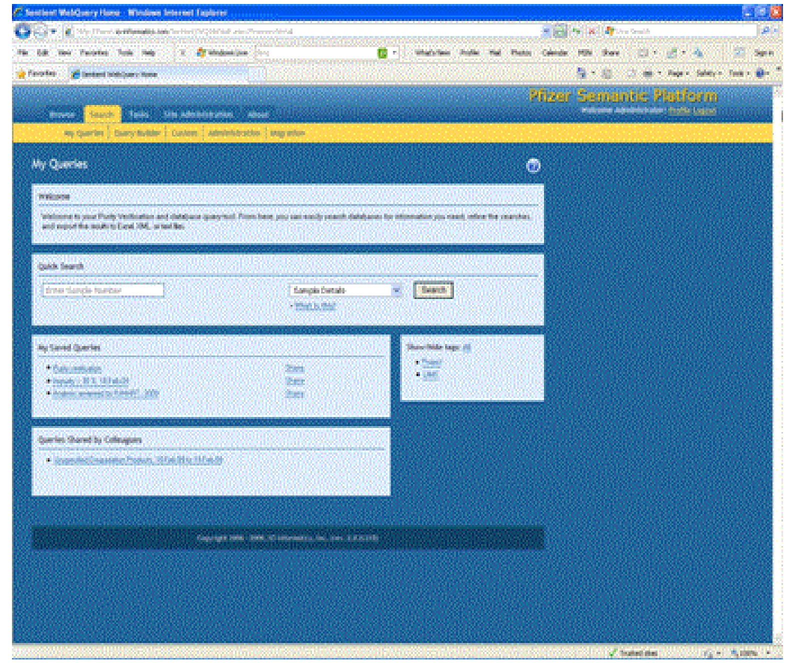
- **Challenges**
  - No common identifiers between LIMS and Empower
  - Limited data stored in Empower (Historical data archived to storage)

# Business Problem #1

IO Informatics Sentient KE & WQ Server
Franz AllegroGraph Database



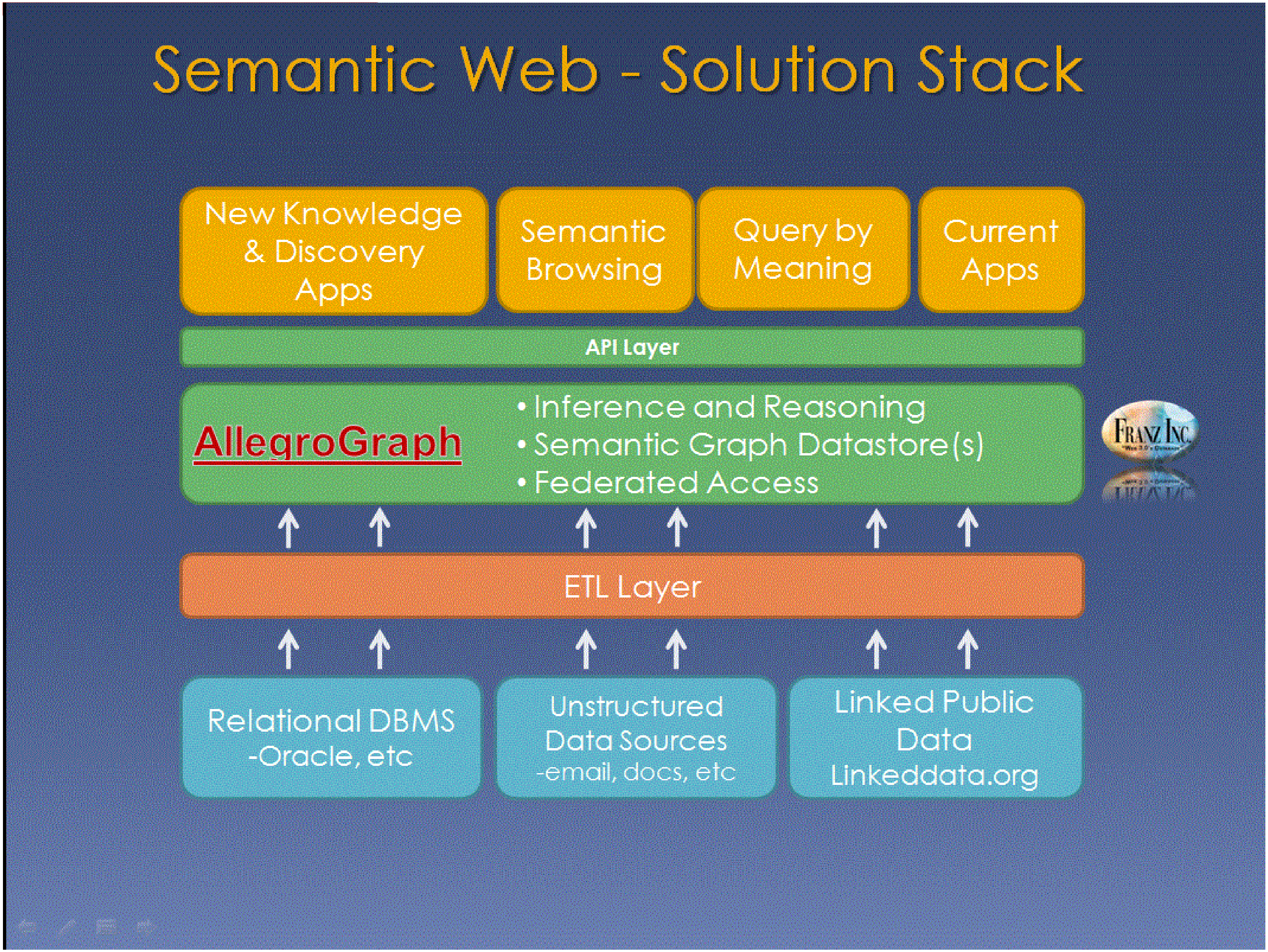

*Imprecise Connections…*



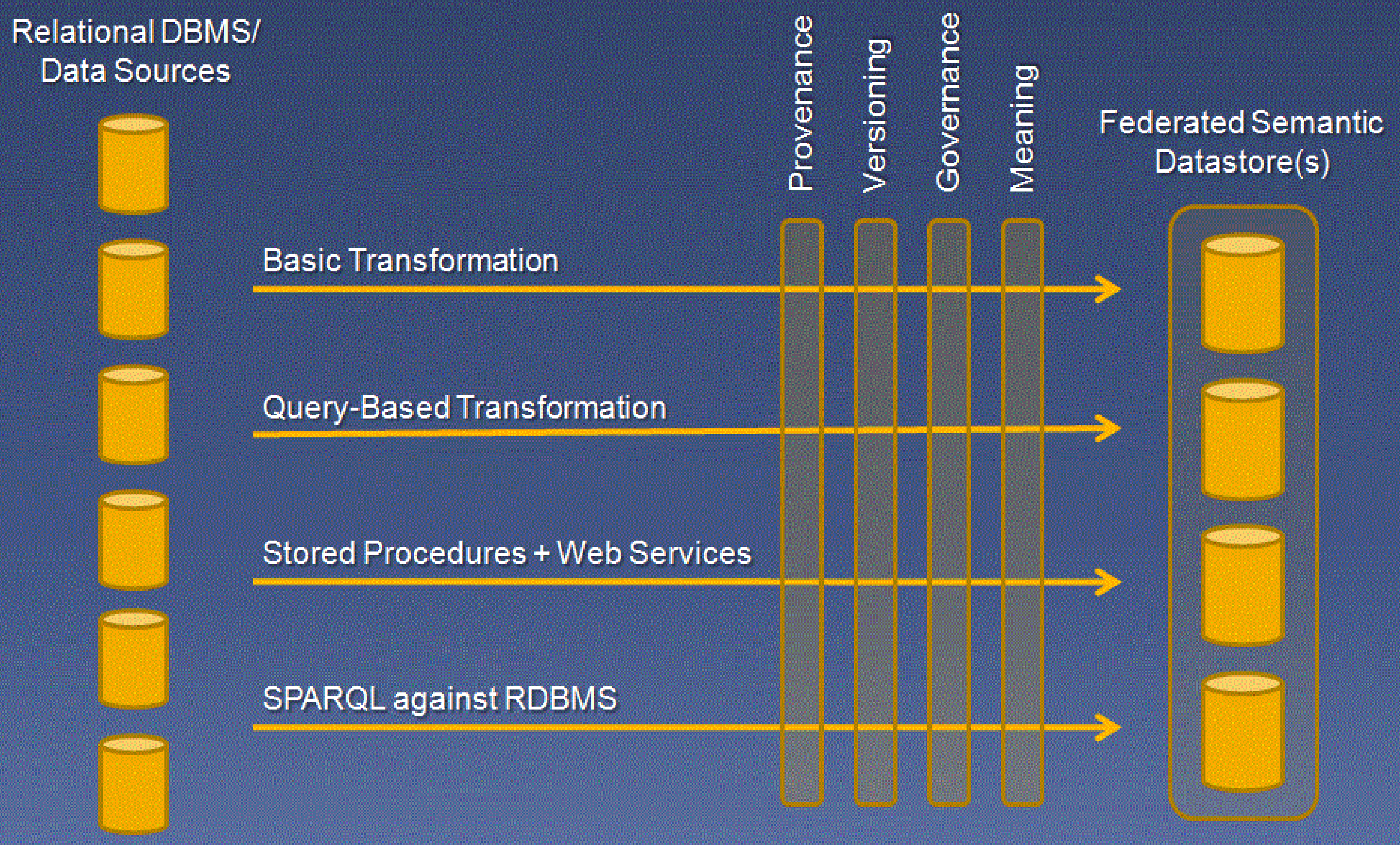

1. Data from LIMS and CDS are transformed via ontologies in context of *multiple imprecise connections.*

2. SPARQL queries across linked
data provide rich pattern-matching capabilities

3. CDS datasets are identified that provide ranked matching to LIMS Reports, for reproducible report verification


Semantic Web - Solution Stack
New Knowledge & Discovery Apps
Semantic Browsing
Query by Meaning
Current Apps
API Layer
AllegroGraph
• Inference and Reasoning
• Semantic Graph Datastore(s)
• Federated Access
Franz Inc.
ETL Layer
Relational DBMS -Oracle, etc
Unstructured Data Sources -email, docs, etc
Linked Public Data Linkeddata.org

# ETL Overview

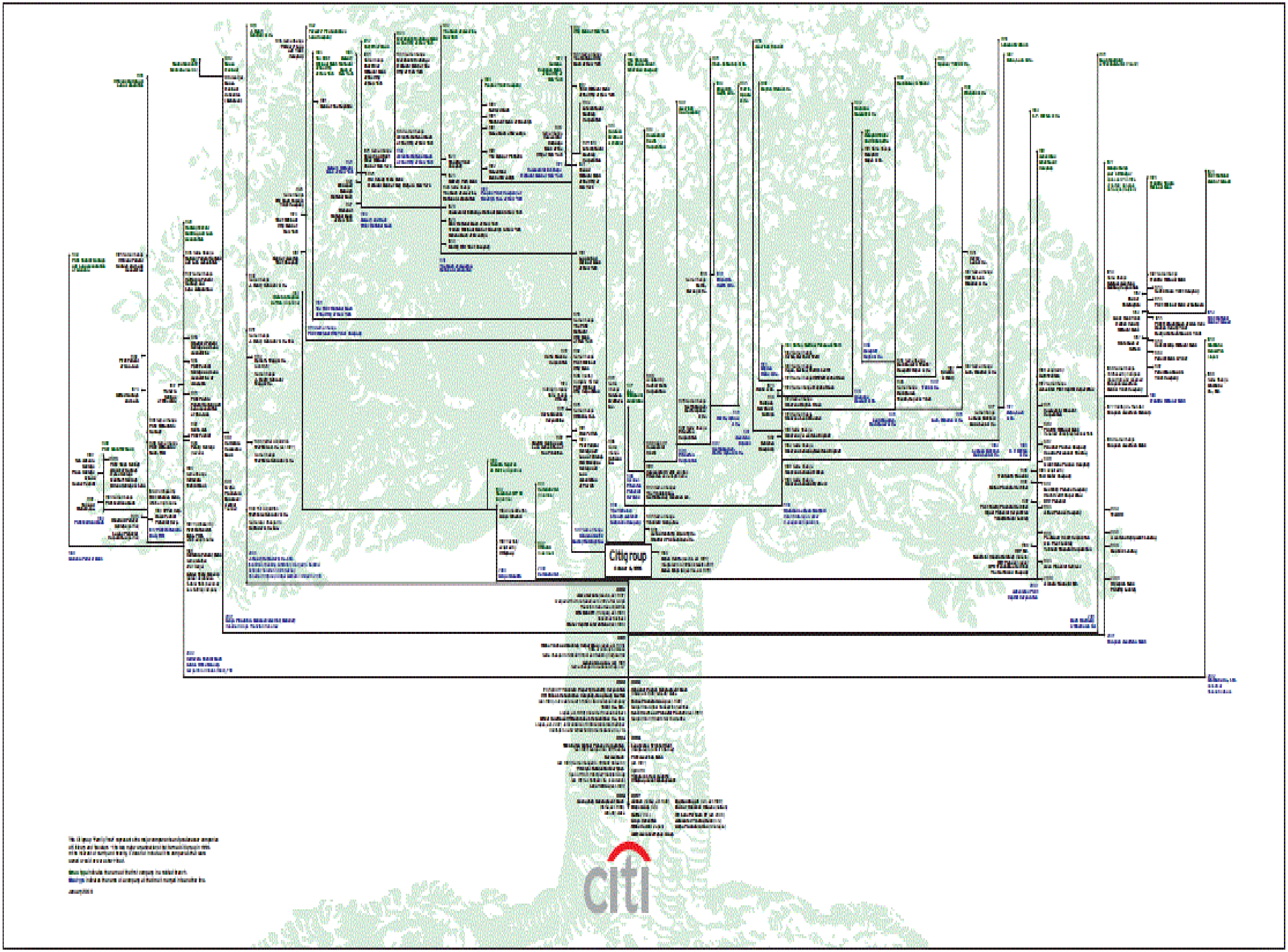


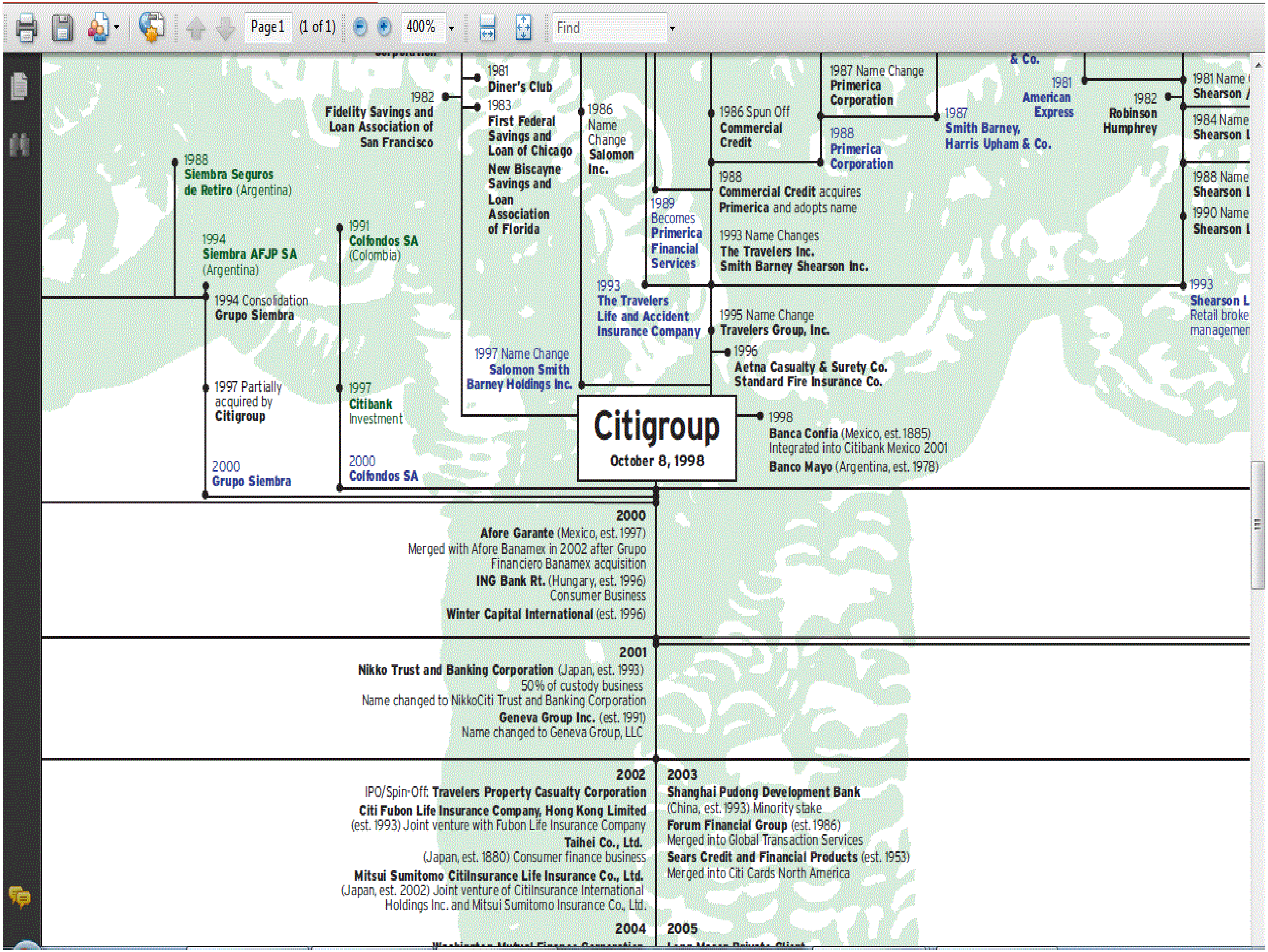
# A Common Pattern

- You have multiple Business Units (Hardware, Software, Services, Applications) that sell all to the same customers
- Each BU ‘result responsible’, so has most efficient set of databases to support own business
  - Customers, contracts, software/hardware versioning, configurations, inventory.
- Only few cross company databases:
  - ERP for accounting and to track sales
  - Customer Care and Trouble Tickets Databases
  - SLA

# Common problems

- Same customer might be in 40 different databases with different customer contacts and account managers, different location addresses and billing addresses.
- Same hardware and software product referenced in many databases, sometimes with different names
- Customers use collections of hardware and software products with different configuration (parameters)
- <u>Inventories</u> discoupled from <u>bill of materials</u> discoupled from <u>customer demand</u> discoupled from <u>problem tickets</u> discoupled from <u>SLA contracts.</u>


Citigroup

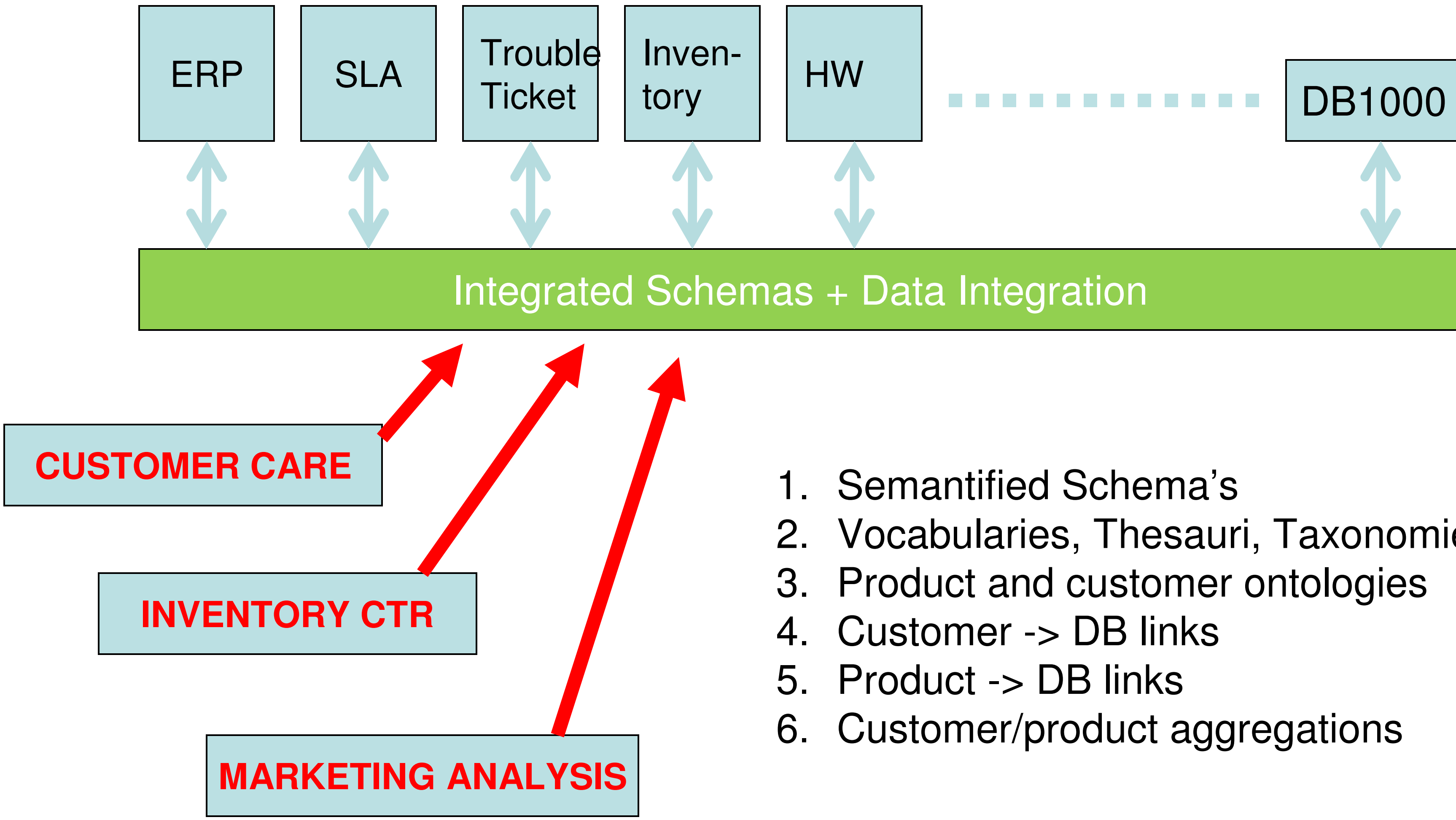


Page 1
(1 of 1)
400%
Find
1981
Diner's Club
1982
Fidelity Savings and Loan Association of San Francisco
1983
First Federal Savings and Loan of Chicago
New Biscayne Savings and Loan Association of Florida
1986
Name Change
Salomon Inc.
1987 Name Change
Primerica Corporation
1986 Spun Off
Commercial Credit
1988
Primerica Corporation
& Co.
1981
American Express
1987
Smith Barney, Harris Upham & Co.
1982
Robinson Humphrey
1981 Name
Shearson
1984 Name
Shearson
1988 Name
Shearson
1990 Name
Shearson
1988
Siembra Seguros de Retiro (Argentina)
1988
Commercial Credit acquires Primerica and adopts name
1989
Becomes
Primerica Financial Services
1993 Name Changes
The Travelers Inc.
Smith Barney Shearson Inc.
1994
Siembra AFJP SA
(Argentina)
1991
Colfondos SA
(Colombia)
1994 Consolidation
Grupo Siembra
1993
The Travelers Life and Accident Insurance Company
1993
Shearson L
Retail broke
managemen
1995 Name Change
Travelers Group, Inc.
1996
Aetna Casualty & Surety Co.
Standard Fire Insurance Co.
1997 Name Change
Salomon Smith Barney Holdings Inc.
1997 Partially acquired by
Citigroup
1997
Citibank
Investment
Citigroup
October 8, 1998
1998
Banca Confia (Mexico, est. 1885)
Integrated into Citibank Mexico 2001
Banco Mayo (Argentina, est. 1978)
2000
Grupo Siembra
2000
Colfondos SA
2000
Afore Garante (Mexico, est. 1997)
Merged with Afore Banamex in 2002 after Grupo Financiero Banamex acquisition
ING Bank Rt. (Hungary, est. 1996)
Consumer Business
Winter Capital International (est. 1996)
2001
Nikko Trust and Banking Corporation (Japan, est. 1993)
50% of custody business
Name changed to NikkoCiti Trust and Banking Corporation
Geneva Group Inc. (est. 1991)
Name changed to Geneva Group, LLC
2002
IPO/Spin-Off: Travelers Property Casualty Corporation
Citi Fubon Life Insurance Company, Hong Kong Limited
(est. 1993) Joint venture with Fubon Life Insurance Company
Taihei Co., Ltd.
(Japan, est. 1880) Consumer finance business
Mitsui Sumitomo CitiInsurance Life Insurance Co., Ltd.
(Japan, est. 2002) Joint venture of CitiInsurance International Holdings Inc. and Mitsui Sumitomo Insurance Co., Ltd.
2003
Shanghai Pudong Development Bank
(China, est. 1993) Minority stake
Forum Financial Group (est. 1986)
Merged into Global Transaction Services
Sears Credit and Financial Products (est. 1953)
Merged into Citi Cards North America
2004
2005

# Impossible question?

- CFO Citigroup: how much do I spend in total with you?
  – yes, he has the same problem ☺
- Sales person: I'm going to sell this video equipment to this company, the customer already has this software/hardware/services configuration, can we expect problems
  – well, apres nous le deluge ☺
- How much do I have to keep in stock given the current rate of problems and the customers that have this in their configuration.
  – Currently we keep *10, just to be sure ☹

# BC




1. Semantified Schema's
2. Vocabularies, Thesauri, Taxonomies
3. Product and customer ontologies
4. Customer -> DB links
5. Product -> DB links
6. Customer/product aggregations

# Traditional Approach: Top Down

- Master Data Management

- Virtual or Federated Database Management

- Think it all out beforehand,
- Heavy Weight,
- Changes are very costly

# Semantic Tech Approach: Bottom up

- Use vocabularies, thesauruses, taxonomies, ontologies
- Translate data into triple stores
- Or query original DB with SPARQL

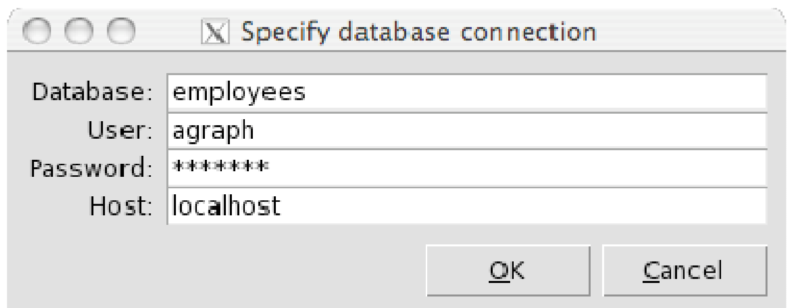
- Lazy, Late binding
- Organic, Evolving
- Very flexible
- Better suited to ad hoc

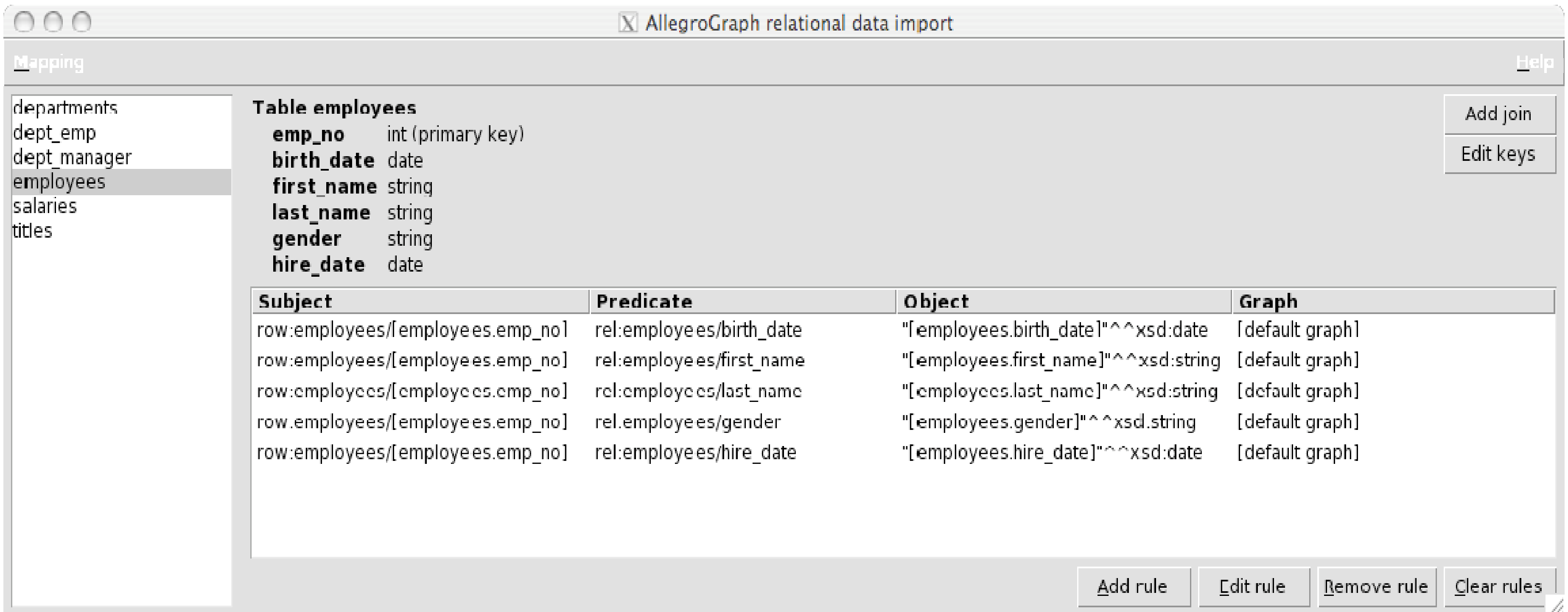
# Step 1: Vocabularies, Thesauri, Taxonomy, Ontologies

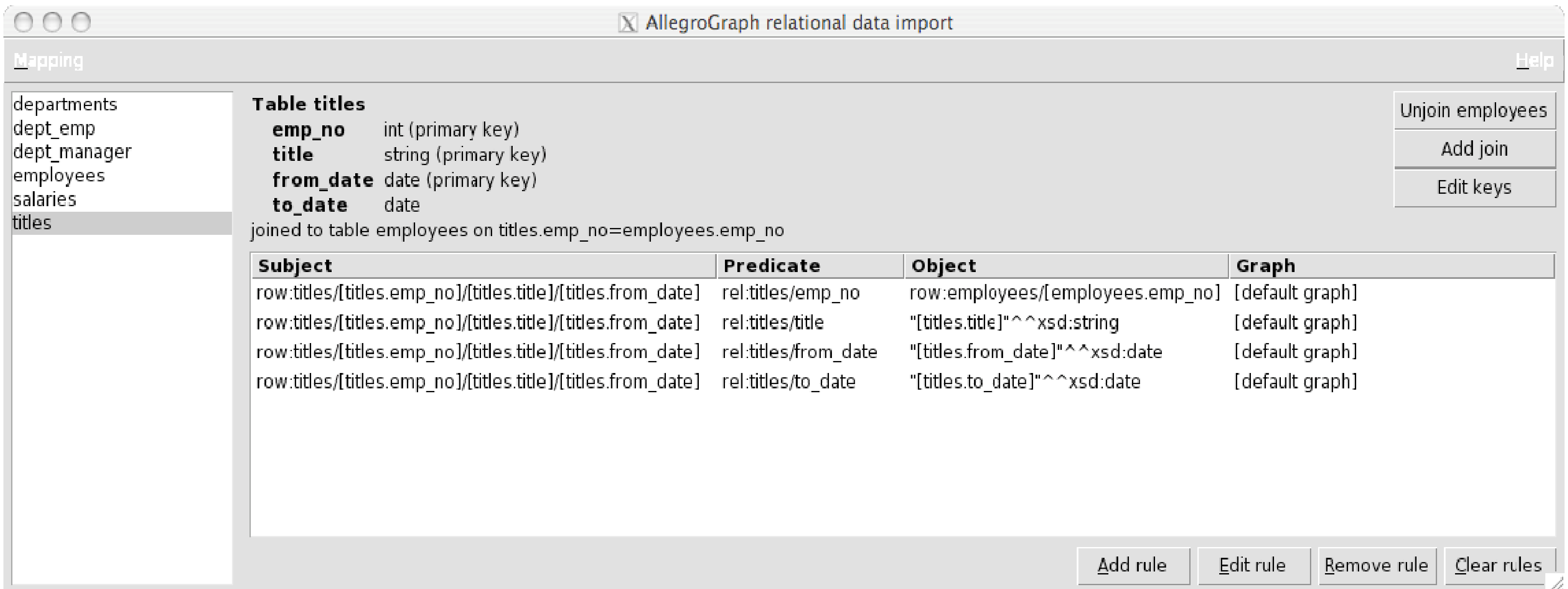
- Vocabularies : the heart of linking
  - `bc:Citi rdf:type bc:VocabularyEntry`
- Thesauri: linking variants to Vocabulary
  - `bc:Citi bc:hasAlternativeName ‘Citi Group’`
- Taxonomy: finding the hierarchy in your data
  - `bc:Banamex bc:part of bc:Citi`
- Ontology: types, subtypes, constraints
  - `bc:Citi rdf:type bc:Bank`
  - `bc:Bank rdf:type owl:Class`
  - `bc:Bank rdfs:subClassOf bc:Company`
  - `bc:Company rdfs:subClassOf bc:Organization`

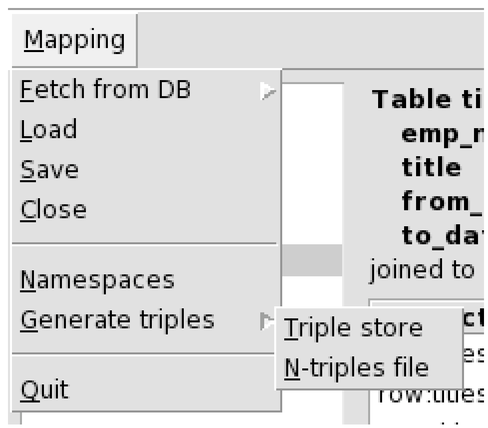
# Step 2: Schema Spaces

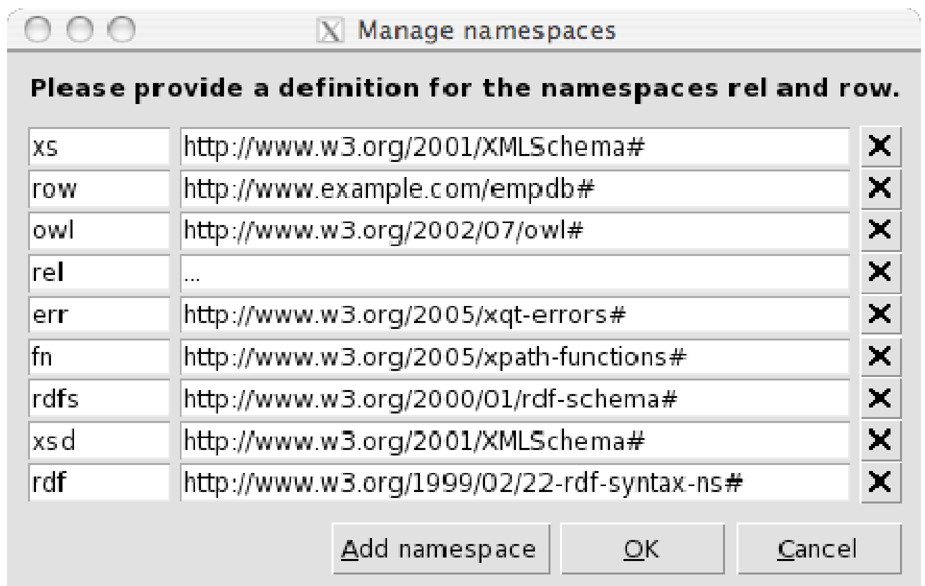
- Create Schema Connection Spaces
  - Take original RDB schemas and syntactically transform to RDF and RDFS
  - `bc:customer1 rdf:type bc:table`
  - `bc:customerID1 rdf:type bc:columnName`
  - `Bc:customerID1  bc:dataType bc:long`
  - Annotate with origin
  - `bc:customer1 bc:fromDB bc:ERP1`
  - Annotate with connections to other schema
  - `bc:customer 1 bc:relatesTo bc:customer2`

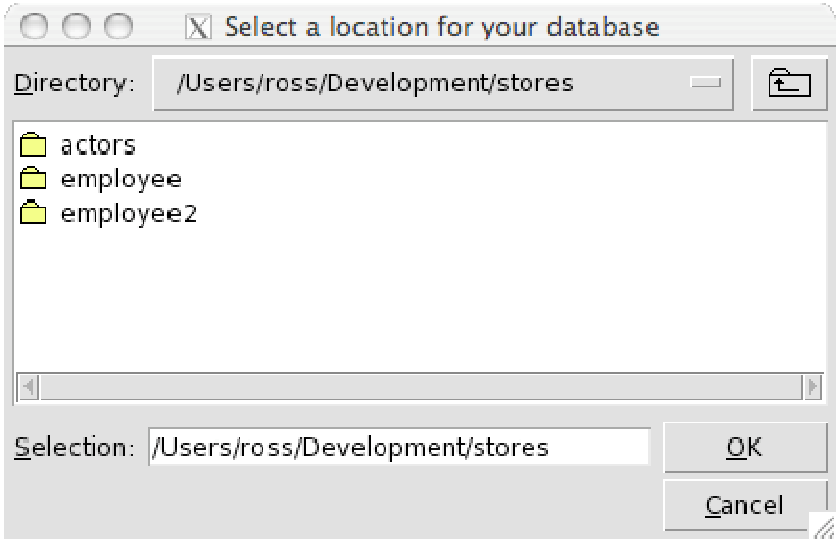

Specify database connection
Database:
employees
User:
agraph
Password:
*******
Host:
localhost
OK
Cancel

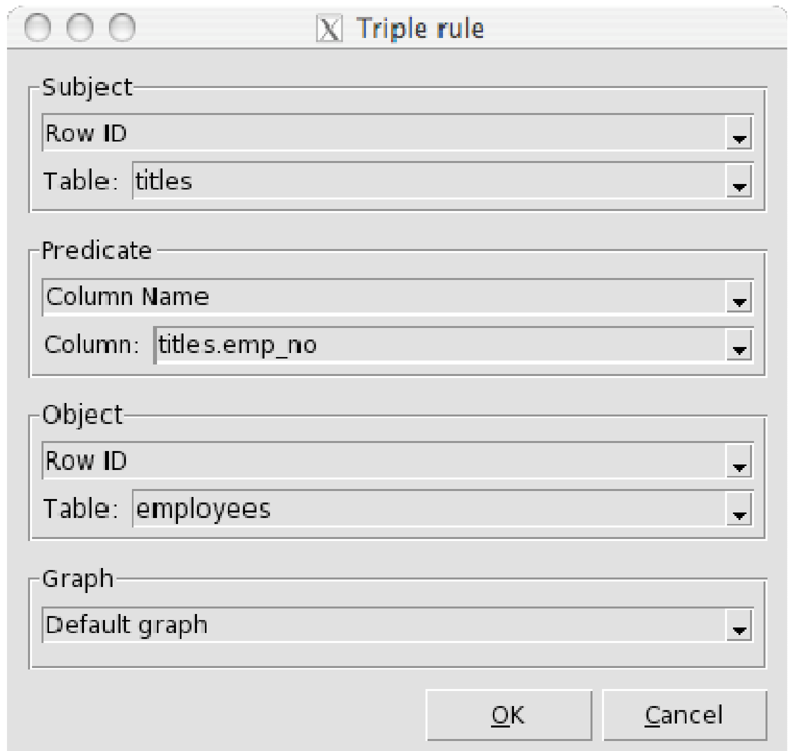

AllegroGraph relational data import
Mapping
Help
departments
dept_emp
dept_manager
employees
salaries
titles
Table employees
emp_no int (primary key)
birth_date date
first_name string
last_name string
gender string
hire_date date
Add join
Edit keys
Subject
Predicate
Object
Graph
row:employees/[employees.emp_no] rel:employees/birth_date "[employees.birth_date]"^^xsd:date [default graph]
row:employees/[employees.emp_no] rel:employees/first_name "[employees.first_name]"^^xsd:string [default graph]
row:employees/[employees.emp_no] rel:employees/last_name "[employees.last_name]"^^xsd:string [default graph]
row:employees/[employees.emp_no] rel:employees/gender "[employees.gender]"^^xsd:string [default graph]
row:employees/[employees.emp_no] rel:employees/hire_date "[employees.hire_date]"^^xsd:date [default graph]
Add rule
Edit rule
Remove rule
Clear rules


AllegroGraph relational data import
Mapping
Help
departments
dept_emp
dept_manager
employees
salaries
titles
Table titles
emp_no int (primary key)
title string (primary key)
from_date date (primary key)
to_date date
joined to table employees on titles.emp_no=employees.emp_no
Unjoin employees
Add join
Edit keys
Subject
Predicate
Object
Graph
row:titles/[titles.emp_no]/[titles.title]/[titles.from_date] rel:titles/emp_no row:employees/[employees.emp_no] [default graph]
row:titles/[titles.emp_no]/[titles.title]/[titles.from_date] rel:titles/title "[titles.title]"^^xsd:string [default graph]
row:titles/[titles.emp_no]/[titles.title]/[titles.from_date] rel:titles/from_date "[titles.from_date]"^^xsd:date [default graph]
row:titles/[titles.emp_no]/[titles.title]/[titles.from_date] rel:titles/to_date "[titles.to_date]"^^xsd:date [default graph]
Add rule
Edit rule
Remove rule
Clear rules

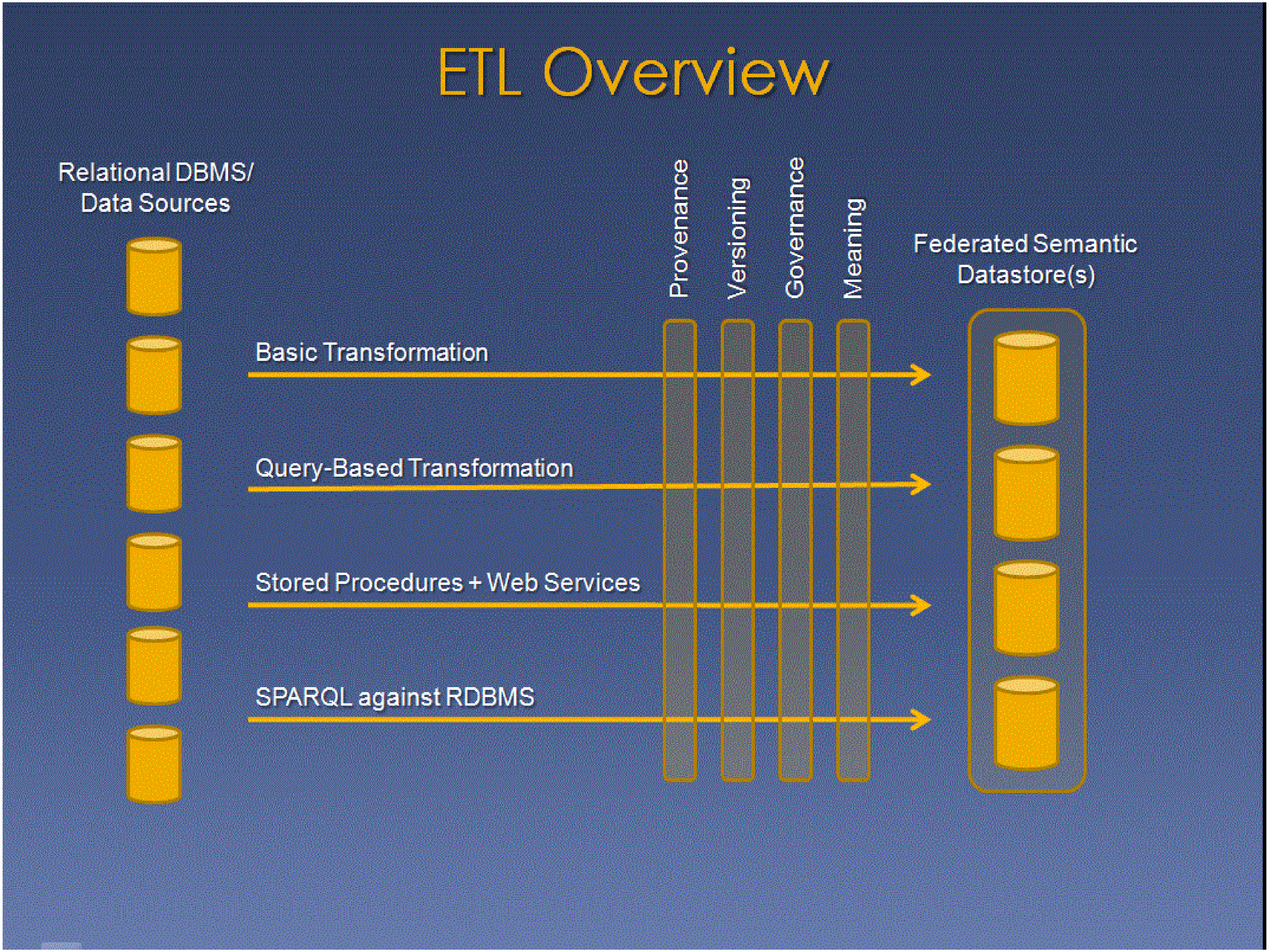

Mapping
Fetch from DB
Load
Save
Close
Namespaces
Generate triples
Quit
Triple store
N-triples file
joined to


Manage namespaces
Please provide a definition for the namespaces rel and row.
xs http://www.w3.org/2001/XMLSchema#
row http://www.example.com/empdb#
owl http://www.w3.org/2002/07/owl#
rel ...
err http://www.w3.org/2005/xqt-errors#
fn http://www.w3.org/2005/xpath-functions#
rdfs http://www.w3.org/2000/01/rdf-schema#
xsd http://www.w3.org/2001/XMLSchema#
rdf http://www.w3.org/1999/02/22-rdf-syntax-ns#
Add namespace
OK
Cancel


Select a location for your database
Directory:
/Users/ross/Development/stores
actors
employee
employee2
Selection:
/Users/ross/Development/stores
OK
Cancel


Triple rule
Subject
Row ID
Table: titles
Predicate
Column Name
Column: titles.emp_no
Object
Row ID
Table: employees
Graph
Default graph
OK
Cancel

# Step 3: RDFy data


ETL Overview
Relational DBMS/
Data Sources
Provenance
Versioning
Governance
Meaning
Federated Semantic
Datastore(s)
Basic Transformation
Query-Based Transformation
Stored Procedures + Web Services
SPARQL against RDBMS

# Step 4: match entities

Entity Resolution

- Is this the same address
- Find same products
- Is this the same company
- Is this the same person

# Step 5: inverted database

- `bc:Citi hasPart bc:Banamex`
- `bc:Banamex  bc:inDB  bc:ERP/customer/name`

# Thanks!