University
of Glasgow

## Use case

- Access and combine data from multiple sources
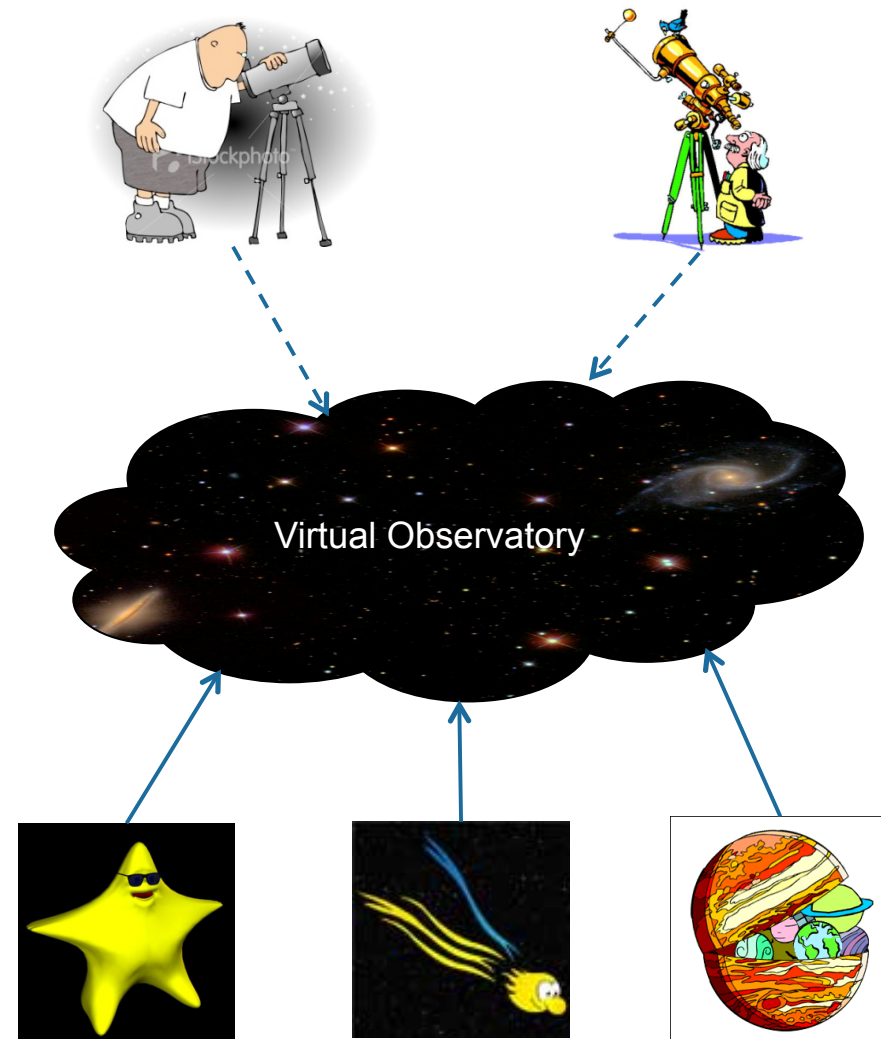- Pose one query

## Issues

Archives exists

- Mostly relational
- Bespoke schemas

Relating data is hard

- Matching columns
- Matching types
- Maintaining semantics
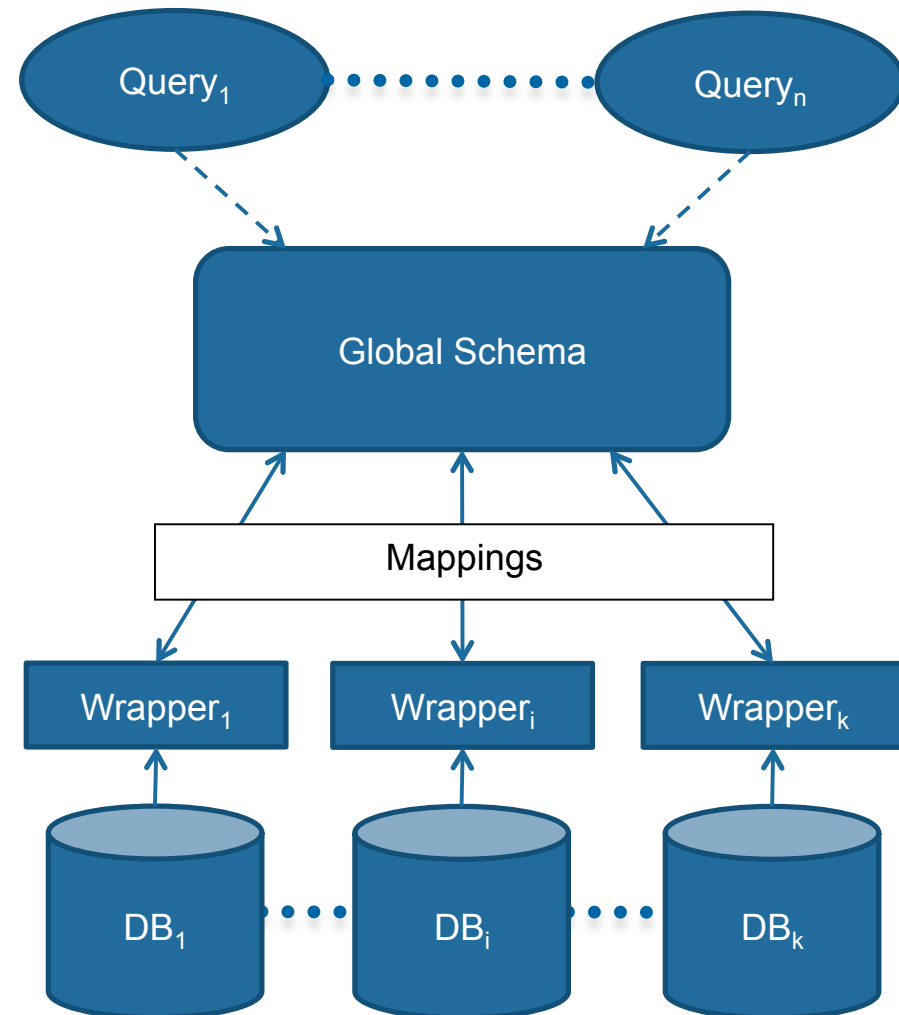
Virtual Observatory

## Data Integration

- Autonomous Sources
  - Maintain own schema
- One Global Schema
  - Pre-agreed by all
  - Limits available data
- Mappings are relational views
  - Global as View (GAV)
  - Local as View (LAV)

## Problem
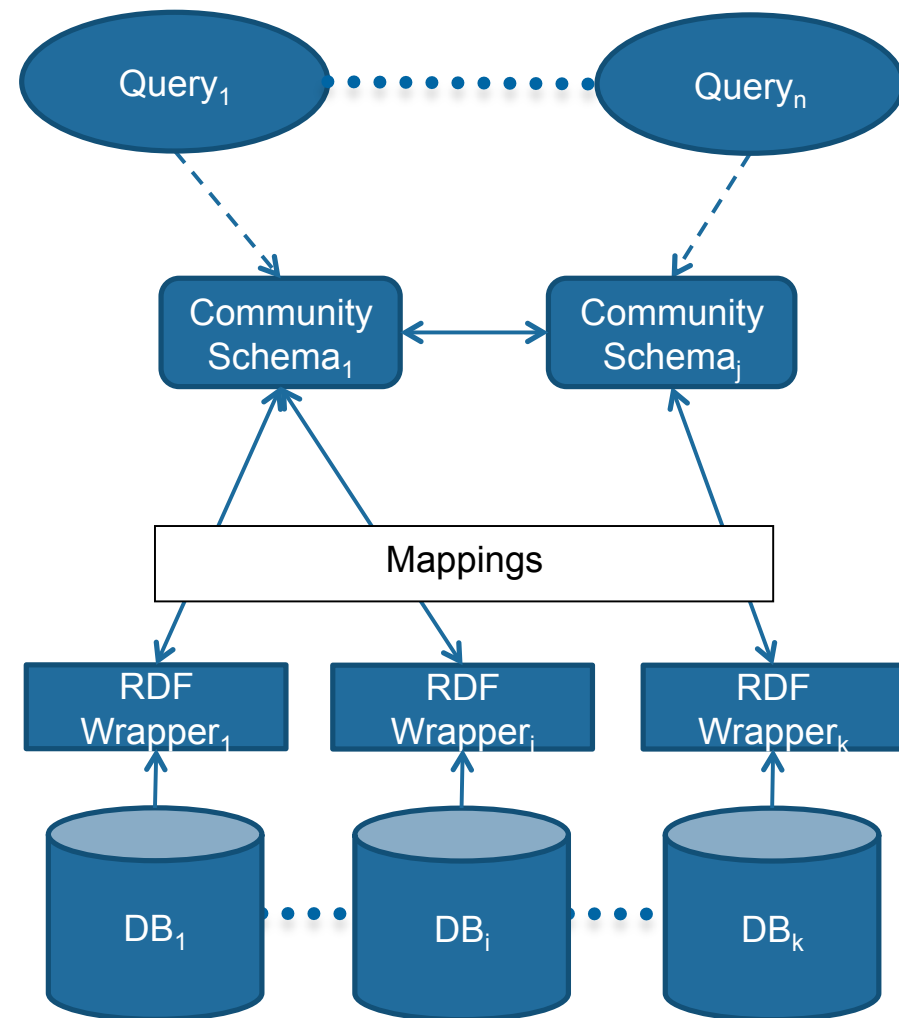
No global schema in astronomy

## Use RDF

- Expose source as RDF model
- Allow multiple *"access"* models
- Semantic mappings between models
- Query using SPARQL

## Two approaches

- Replicate data in RDF Store
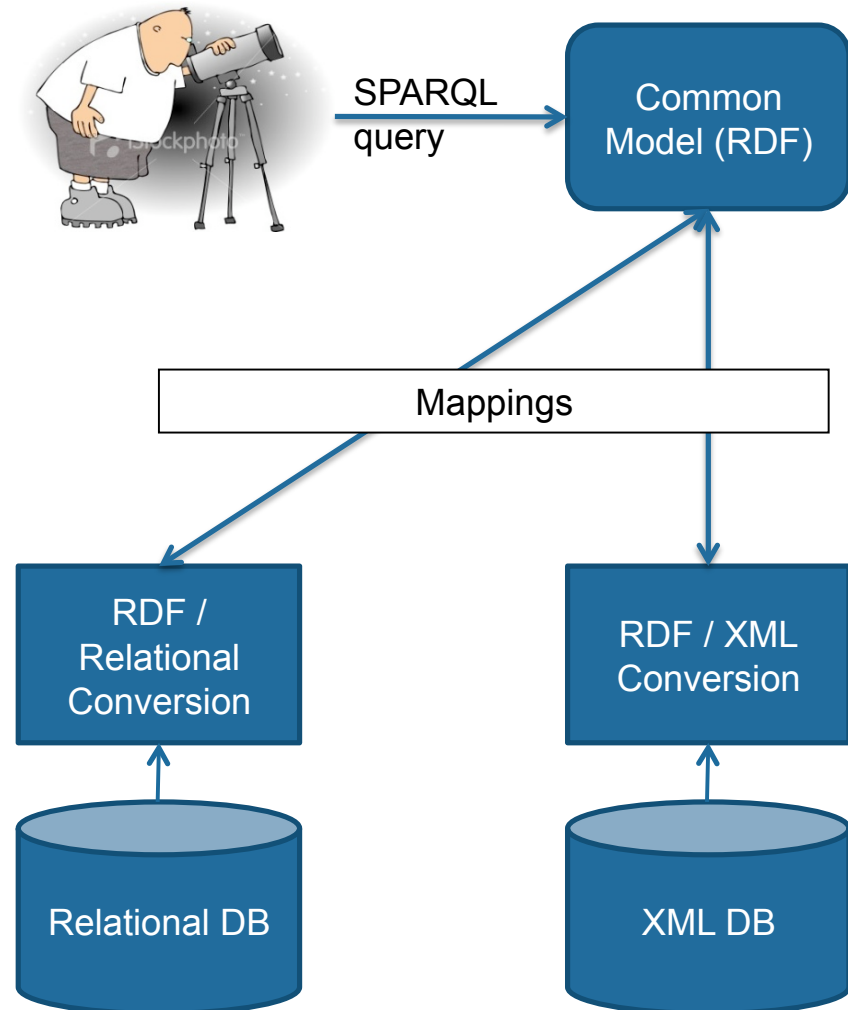  - Consistency issues
- On-the-fly translation
  - SPARQL to SQL
  - Relational data to RDF

**Is it viable to perform *on-the-fly* conversions from existing science archives to RDF to facilitate data access from a data model that a scientist is familiar with?**

SPARQL query → Common Model (RDF)

Mappings

RDF / Relational Conversion

RDF / XML Conversion

Relational DB

XML DB

## SuperCOSMOS Science Archive (SSA)

Data extracted from scans of Schmidt plates

Stored in a relational database

About 4TB of data, detailing 6.4 billion objects

Fairly typical of astronomical data archives

## Schema designed using 20 real queries

## Personal version contains

Data for a specific region of the sky

About 0.1% of the data, ~500MB

**Query 5**

**Find the positions and (B,R,I) magnitudes of all star-like objects within delta mag of 0.2 of the colours of a quasar of redshift 2.5 < z < 3.5**

```
SELECT TOP 30 ra, dec,
sCorMagB, sCorMagR2,
sCorMagI

FROM ReliableStars

WHERE (sCorMagB-
sCorMagR2 BETWEEN 0.05
AND 0.80) AND
(sCorMagR2-sCorMagI
BETWEEN -0.17 AND 0.64)
```

| Query Feature | Query Numbers |
| --- | --- |
| Arithmetic in body | 1-5, 7, 9, 12, 13, 15-20 |
| Arithmetic in head | 7-9, 12, 13 |
| Ordering | 1-8, 10-17, 19, 20 |
| Joins (including self-joins) | 12-17, 19 |
| Range functions (e.g. Between, ABS) | 2, 3, 5, 8, 12, 13, 15, 17-20 |
| Aggregate functions (including Group By) | 7-9, 18 |
| Math functions (e.g. power, log, root) | 4, 9, 16 |
| Trigonometry functions | 8, 12 |
| Negated sub-query | 18, 20 |
| Type casting (e.g. Radians to degrees) | 7, 8, 12 |
| Server functions | 10, 11 |

## Features

- Select-project-join
- Arithmetic in body
- Conjunction and disjunction
- Ordering
- String matching
- External function calls
  (extension mechanism)

## Limitations

- Range shorthands
- Arithmetic in head
- Math functions
- Trigonometry functions
- Sub queries
- Aggregate functions
- Casting

| Query Feature | Query Numbers | |
|---|---|---|
| Arithmetic in body | 1-5, 7, 9, 12, 13, 15-20 | ✔ |
| Arithmetic in head | 7-9, 12, 13 | ✘ |
| Ordering | 1-8, 10-17, 19, 20 | ✔ |
| Joins (including self-joins) | 12-17, 19 | ✔ |
| Range functions (e.g. Between, ABS) | 2, 3, 5, 8, 12, 13, 15, 17-20 | ✔ |
| Aggregate functions (including Group By) | 7-9, 18 | ✘ |
| Math functions (e.g. power, log, root) | 4, 9, 16 | ✘ |
| Trigonometry functions | 8, 12 | ✘ |
| Negated sub-query | 18, 20 | ✘ |
| Type casting (e.g. radians to degrees) | 7, 8, 12 | ✘ |
| Server functions | 10, 11 | ✘ |

**Expressible queries: 1, 2, 3, 5, 6, ~~14, 15, 17, 19~~**

## Machine

Quad Core Intel Xeon 2.4GHz 64 bit processor

4GB RAM

100 GB Disc

Linux

Java 1.6

## Software

Database

- MySQL 5.1.25

Triple Stores

- Jena 2.5.6 with SDB 1.1

- Sesame 2.1.3

RDB2RDF Convertors

- D2RQ 0.5.2

- SquirrelRDF 0.1

Only 5 queries completed within 2 hours

## Query Execution (ms)

| Query | 1 | 2 | 3 | 5 | 6 |
|---|---|---|---|---|---|
| JDBC/MySQL (Views) | 34 | 38 | 33 | 34 | 1 |
| D2RQ (Views) | 352 | 5,339 | 2,733 | 4,090 | 7,468 |
| D2RQ | 39,374 | | 40,021 | 153,392 | |
| SquirrelRDF (Views) | 613 | 21,492 | 837 | 1,307 | 19,984 |
| Jena SDB (Views) | 3,450 | 485,932 | 7,229 | 17,793 | 372,561 |
| Sesame (Views) | 39 | 83 | 69 | 65 | 56 |
| Sesame | 88 | | 122 | 128 | |

**Exploit query engines and data structure of underlying data sources**

**Aid user query generation by *explaining* source data model in terms of known data model**

**Data extracted in native model**

**Explicator**

| Models | ⟷ | Mappings |

Explain

SQL

XQuery

Relational DB

XML DB
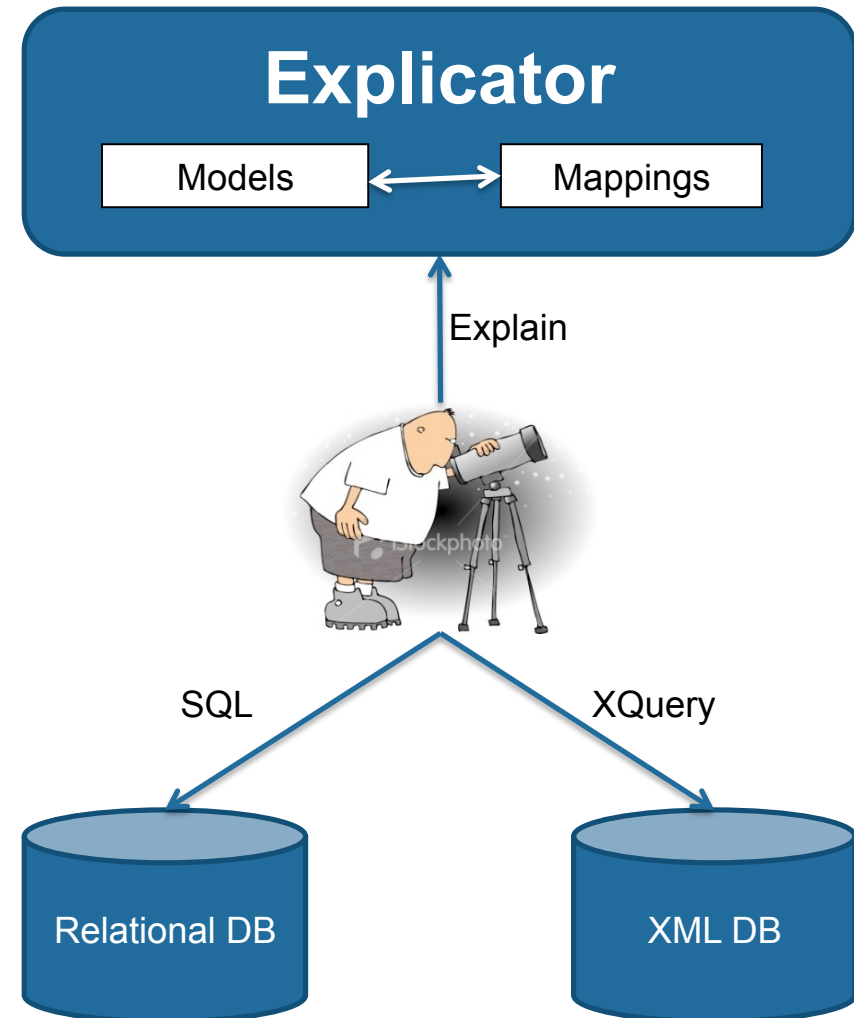
| RDF | Relational |
|---|---|
| Ragged data | Structured data |
| Small to medium data volumes | Large data volumes |
| Reasoning over the data | Extracting specific data |

**SPARQL: Not expressive enough for science**

**Query Converters: Poor performance**

**Exploring a new approach**

RDF to understand data models

Native query engines for data extraction

University
of Glasgow

VIA VERITAS VITA

# PRACTICAL SEMANTIC ASTRONOMY 2009

## 2-5 MARCH 2009

## GLASGOW, UK

seeAlso

mag +3.1

star

isA

isA

A3 dwarf

isA

Megrez

hasA

isA

object

hasStars

partOf

Constellation

Ursa Major

hasName