

# Triple store databases and their role in high throughput, automated extensible data analysis



Jeremy Frey

ACS San Diego March 2005

March 2005

Jeremy Frey

ACS CINF



## Talk: Workflow

- Introduction to the Combechem Project
- Smart Dark Labs
- Semantics & Databases
- RDF and Triple Stores
- Future

March 2005

Jeremy Frey

ACS



## e-Science

'[The Grid] intends to make access to computing power, scientific data repositories and experimental facilities as easy as the Web makes access to information.' Tony Blair, 2002

- What is the web?
- Need to worry about what the data means as well as simple numerical value

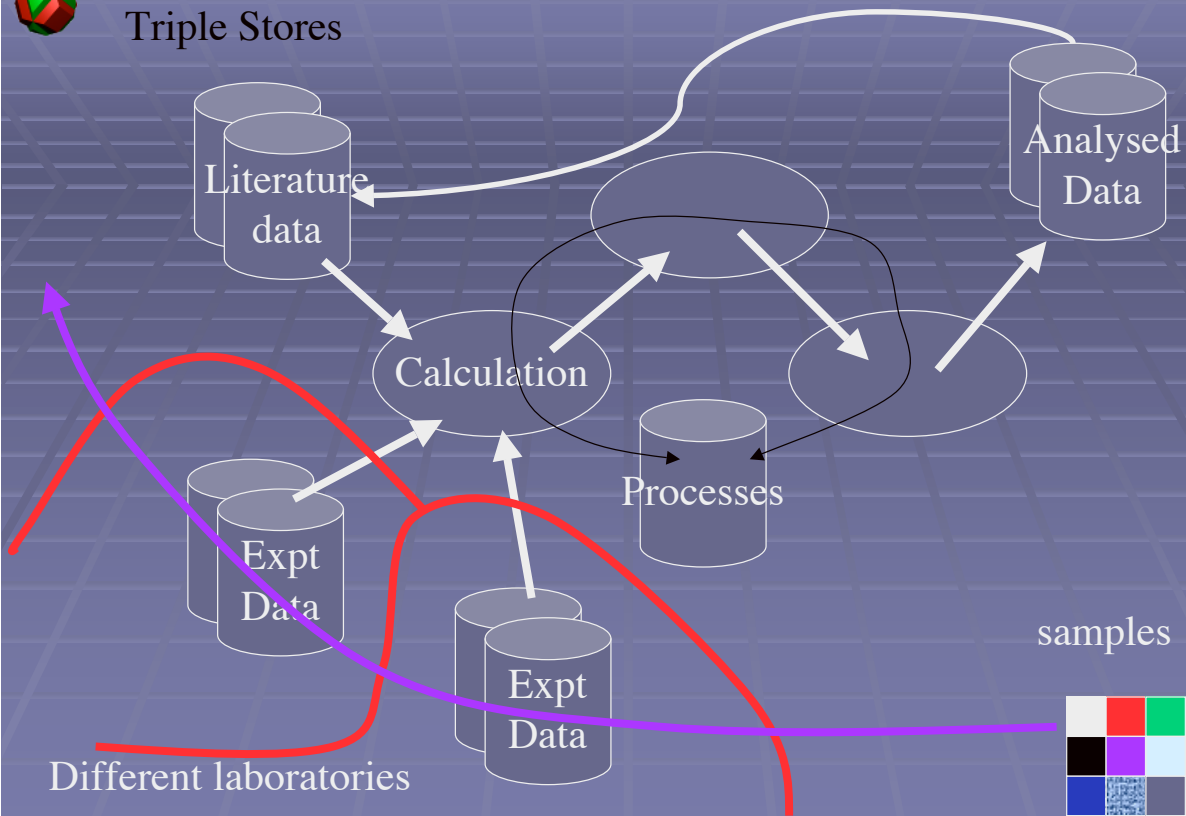


## The CombeChem Project

- The exponential world of combinatorial synthesis and high throughput analysis meets the exponentially growing power of computing
- Provenance:- Trace all the way back from publication to the original data which may be in several different labs
- But then... "Who wants provenance?"  
Bush, Blair The JIC, MI5, CIA & Hutton 2004
- Worked on high-throughput molecular crystallography
- Extend to other labs and computation - need to take a look at the way we store chemical information



# Triple Stores



Different laboratories

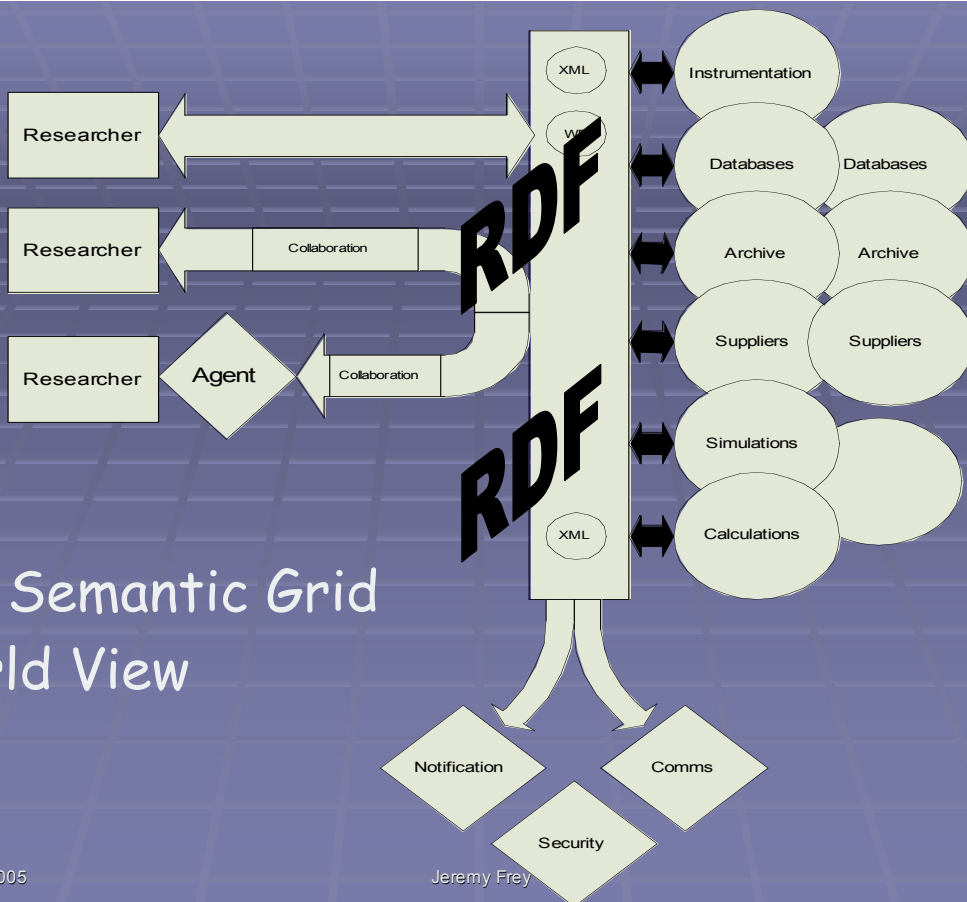
March 2005

Jeremy Frey

ACS



# The Semantic Grid World View



March 2005

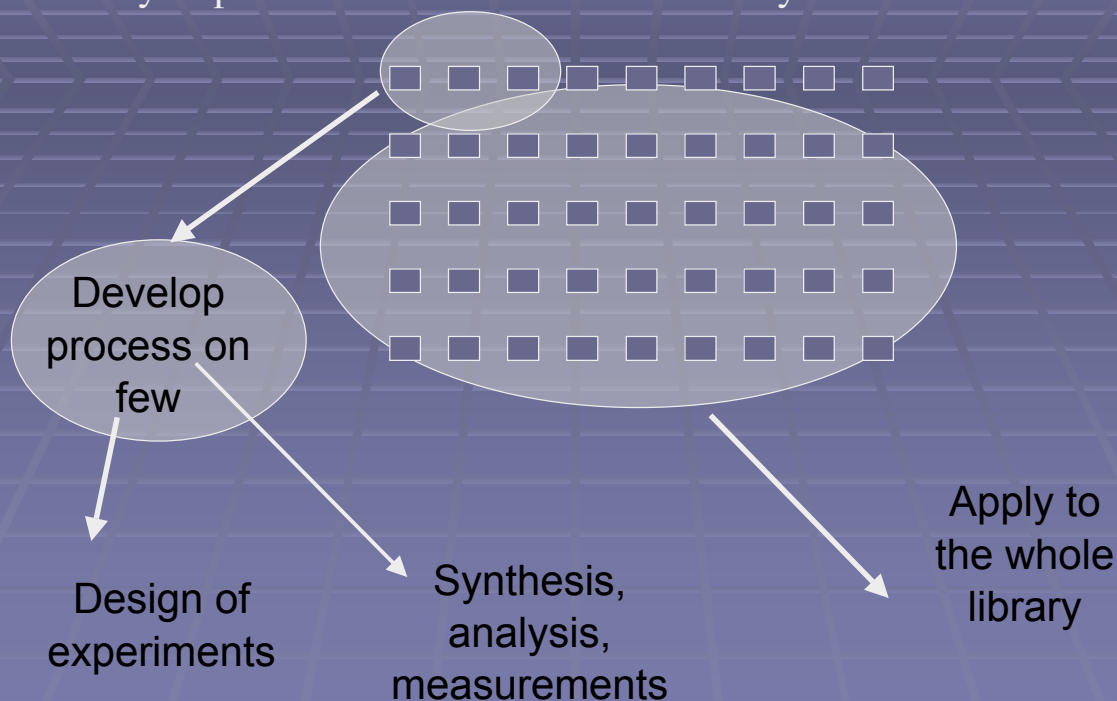
Jeremy Frey

ACS



# High-throughput vs Parallel

May impose a real time constraint in analysis



March 2005

Jeremy Frey

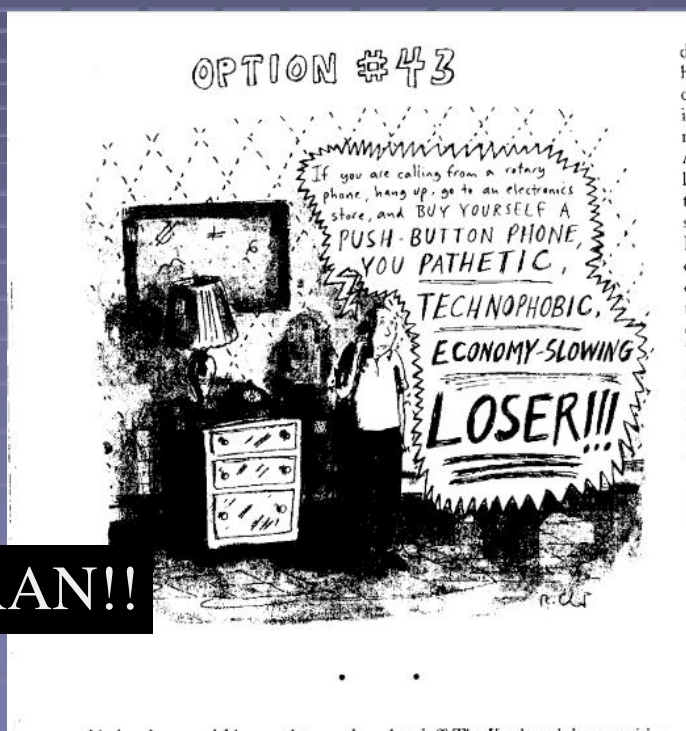
ACS



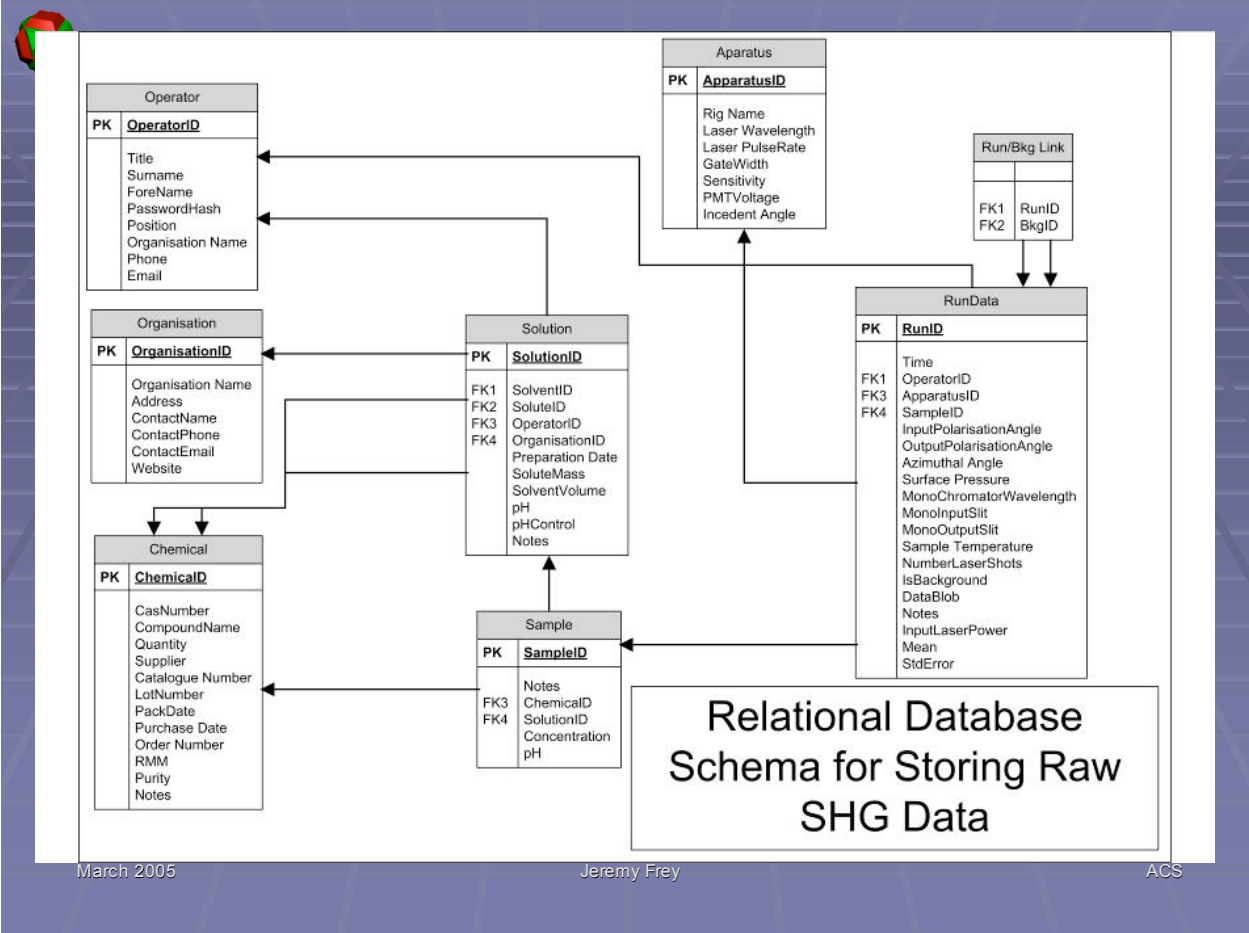
# Chemists and programming

- Many Chemists think that they can program
- So leave the systems to the Chemists

**You still use FORTRAN!!**



March 2005



## Databases - Our experience

- What do you do when the actual users keep changing their mind?
- Is a traditional relational database suitable?
- Danger of re-enforcing scientific bias against relational database for laboratory data.
- Semantic Web (or Grid?) - using RDF



# RDF

- Triples
  - Subject, predicate, object
- Results in a chain of reasoning that can be mapped by generic inference software
- Real power comes from the data becomes self describing
- Together with an Ontology the data becomes 'alive'



# RDF

- Project started in 1997 - W3C standard in 2000
- Aim to enable the Semantic Web
  - But how many are there even though we have had the standard for 4 years
- Enable search engines to collect information from many sources
- At present requires deliberate human controlled efforts
- Advent of a critical mass of metadata facilitates machine-based discovery.



# RDF

- Capable of representing more complex structures than XML which is fundamentally a tree like structure
- Relationships like 'similar to' can be employed
- Not just for web & sharing
- Useful as a local storage and reference system



# Chemical data

- Chemical data
  - Requires extensive annotation
  - Purity, method, accuracy, conditions
- RDF
  - Schema can supply this information
  - Easy to add new information
  - Very flexible
  - Perhaps issues of flexibility over speed



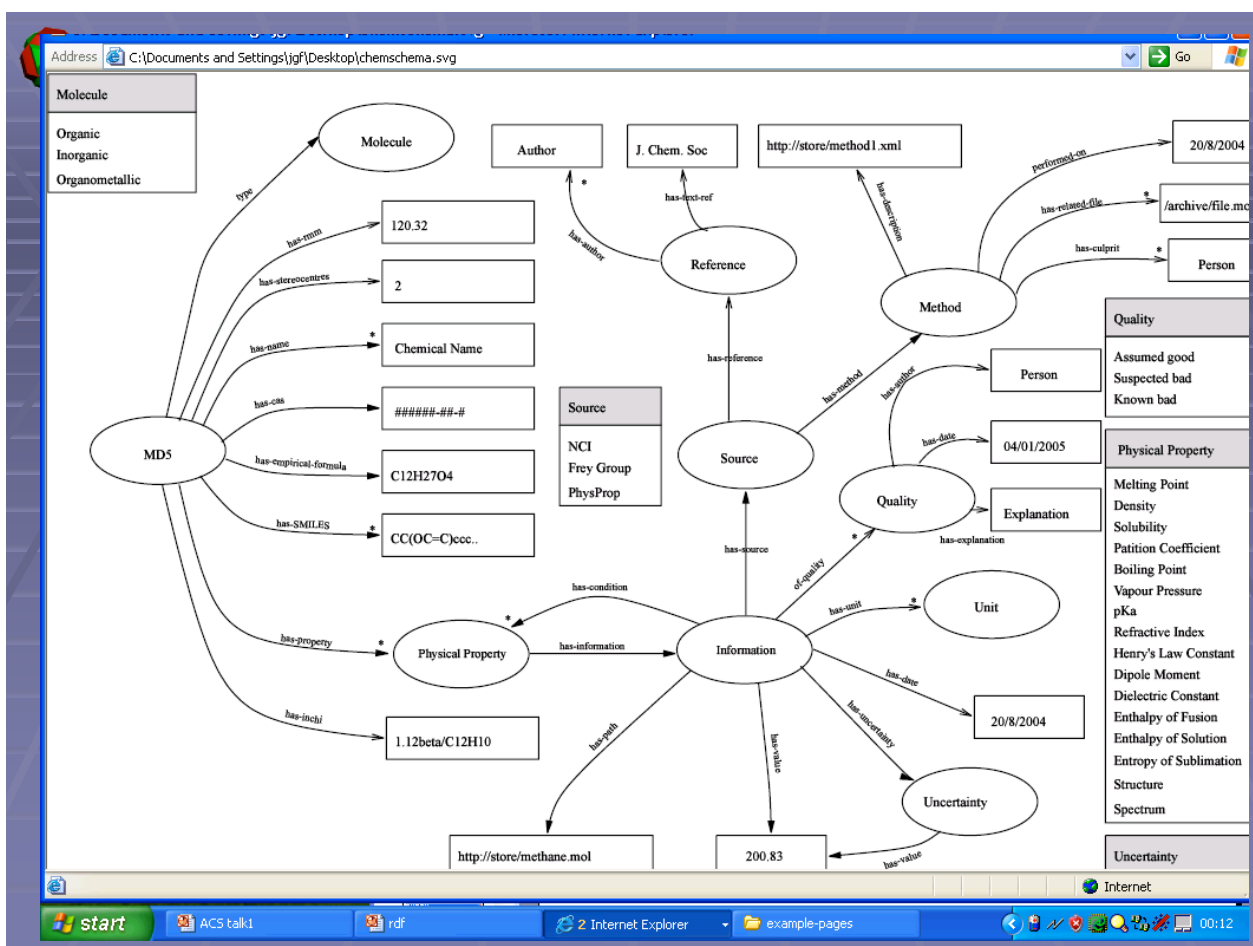
# root

- INChI (International Chemical Identifier)
  - Can be too long so use MD5 hash
- Molecule centric view
  - Full stereochemistry (if known)
  - Phase & Polymorphs
  - Structure
    - Calculated, derived, crystal
- Properties
  - Contains tracking information
  - Uncertainty, Source, Reference

March 2005

Jeremy Frey

ACS







- `<c:OrganicMolecule`  
`rdf:about="file:///storage/ba8efc2ce0edada69d63b02d1b8630c6.rdf">`
- `<c:has-inchi>1.12Beta/C12H13NO2/c1-2-15-8-9-5-6-11(14)12-10(9)4-3-7-`  
`13-12/h1H3,2H2,3-7H,8H2,14H</c:has-inchi>`
- `<c:has-cas>22049-19-0</c:has-cas>`
- `<c:has-empirical-formula>C12H13NO2</c:has-empirical-formula>`
- `<c:has-stereocentres>0</c:has-stereocentres>`
- `<c:has-property`
- `<c:MeltingPoint`
- `<c:has-information`
- `<c:Information`
- `<c:has-value>150</c:has-value>`
- `<c:has-uncertainty`
- `<c:Range`
- `<c:has-value>16</c:has-value>`
- `</c:Range`
- `</c:has-uncertainty`
- `</c:Information`
- `</c:has-information`
- `</c:MeltingPoint`
- `</c:has-property`
- `</c:OrganicMolecule`

## Property in RDF

*Currently testing on  
200,000 compounds but  
about to go up by order of  
magnitude*

*3Store is a scaleable  
solution*

March 2005

Jeremy Frey

ACS



## Schema

- `<rdfs:Class rdf:about="&c;OrganicMolecule">`
- `<rdfs:label>Organic Molecule</rdfs:label>`
- `<rdfs:subClassOf rdf:resource="&c;Molecule" />`
- `</rdfs:Class>`
- `<rdfs:Class rdf:about="&c;PhysicalProperty">`
- `<rdfs:label>Property</rdfs:label>`
- `</rdfs:Class>`
- `<rdfs:Class rdf:about="&c;PartitionCoefficient">`
- `<rdfs:label>Parition Coefficient</rdfs:label>`
- `<rdfs:subClassOf rdf:resource="&c;PhysicalProperty" />`
- `<rdfs:description>Ratio of substance dissolved in octan-1-ol and water`  
`</rdfs:description>`
- `</rdfs:Class>`

This turns out to be a very flexible approach

March 2005

Jeremy Frey

ACS



# Triple Store

- Simply a store for a set of RDF triples
- Collection of triples can then be used for inference as each triple is an assertion
- Several triple stores are available – JENA
- But we require a very large triple store
  - 250,000 molecules some with properties needed 10 million triples
  - 100 million triples to represent the ZINC database

March 2005

Jeremy Frey

ACS

**Jena – A Semantic Web Framework for Java**

Jena is a Java framework for building [Semantic Web](#) applications. It provides a programmatic environment for [RDF](#), [RDFS](#) and [OWL](#), including a rule-based inference engine. Jena is [open source](#) and grown out of work with the [HP Labs Semantic Web Programme](#).

The Jena Framework includes:

- A RDF API
- Reading and writing RDF in RDF/XML, N3 and N-Triples
- An OWL API
- In-memory and persistent storage
- RDQL – a query language for RDF

Support is provided by the [jena-dev](#) mailing list.

**HP Labs**

The HP Labs Semantic Web research group believes that the Semantic Web represents a huge potential technology disrupter, enabling new and more flexible approaches to data integration, web services and knowledge discovery. The group is committed to furthering the Semantic Web vision through:

- standards work via the W3C to support the growth of the Semantic Web;
- tools development to encourage the exploration and exploitation of the Semantic Web by developers;
- core research to help develop the field;
- applications research to demonstrate the value of Semantic Web;
- consultancy within HP to promote its use within the company.

**News:**

- February 2005 - new student intern places available - check "Student Projects" page under "HP Labs Research"
- February 2005 - [HP Labs 0.8.2 released](#)
- January 2005 - [HP Labs 0.8.1 released](#)
- December 2004 - [HP Labs 0.8.0 released](#)
- March 2004 - [Jena 1.1 released](#) to coincide with the [W3C announcement](#) of the new RDF and OWL standards

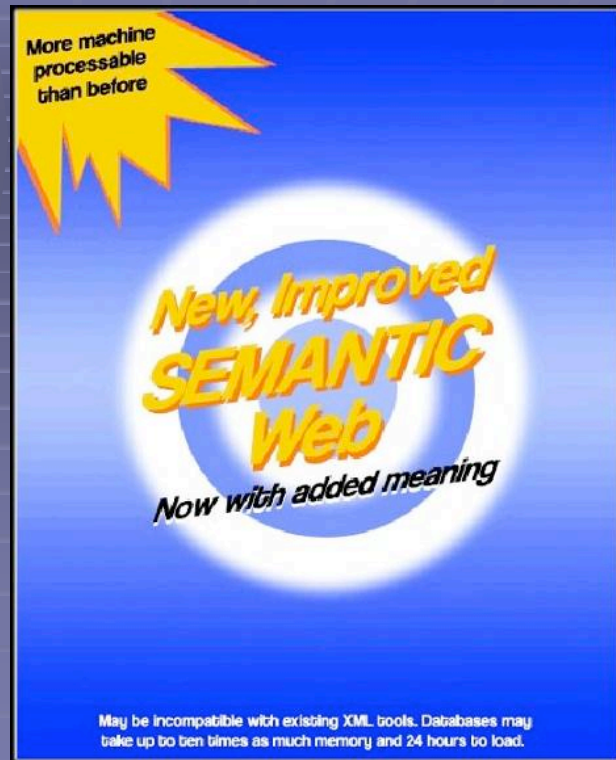
**Not a scalable solution but does allow for more general inference queries**

March 2005



Semantic web  
-But how to store it?

-Knowledge  
Technologies



March 2005

Jeremy Frey

ACS

Let  $\mathbf{A}$ ,  $\mathbf{R}_A$ ,  $\mathbf{R}_D$ , and  $\mathbf{I}$  be pairwise disjoint sets of *concept names*, *abstract role names*, *datatype (or concrete) role names*, and *individual names*. The set of  $SHOIN(\mathbf{D})$ -roles is  $\mathbf{R}_A \cup \{R^- \mid R \in \mathbf{R}_A\} \cup \mathbf{R}_D$ . In order to avoid considering roles such as  $R^{--}$  we will define  $\text{Inv}(R)$  s.t.  $\text{Inv}(R) = R^-$  and  $\text{Inv}(R^+) = R$ . The set of  $SHOIN(\mathbf{D})$ -concepts is the smallest set that can be built using the constructors in Figure 3.

The  $SHOIN(\mathbf{D})$  axioms are listed in Figure 3. (The last axiom in Figure 3 forms an extension of  $SHOIN(\mathbf{D})$  which we call  $SHOIN^+(\mathbf{D})$ , which is used internally in our translation.) A *knowledge base*  $\mathcal{K}$  is a finite set of axioms. We will use  $\sqsubseteq$  to denote the transitive reflexive closure of  $\sqsubseteq$  on roles, i.e., for two roles  $S, R$  in  $\mathcal{K}$ ,  $S \sqsubseteq R$  in  $\mathcal{K}$  if  $S = R$ ,  $S \sqsubseteq R \in \mathcal{K}$ ,  $\text{Inv}(S) \sqsubseteq \text{Inv}(R) \in \mathcal{K}$ , or there exists some role  $Q$  such that  $S \sqsubseteq^* Q$  in  $\mathcal{K}$  and  $Q \sqsubseteq R$  in  $\mathcal{K}$ . A role  $R$  is called *simple* in  $\mathcal{K}$  if for each role  $S$  s.t.  $S \sqsubseteq R$  in  $\mathcal{K}$ ,  $\text{Trans}(S) \notin \mathcal{K}$  and  $\text{Trans}(\text{Inv}(S)) \notin \mathcal{K}$ . To maintain decidability, a knowledge base must have no number restrictions on non-simple roles [11].

The semantics of  $SHOIN^+(\mathbf{D})$  is given by means of an interpretation  $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$  consisting of a non-empty domain  $\Delta^{\mathcal{I}}$ , disjoint from the datatype (or concrete) domain  $\Delta_{\mathbf{D}}^{\mathcal{I}}$ , and a mapping  $\cdot^{\mathcal{I}}$ , which interprets atomic and complex concepts,



AKT - Technologies - 3store from The University of Southampton - Microsoft Internet Explorer

Address <http://www.aktors.org/technologies/3store/>

ADVANCED KNOWLEDGE  
**AKT**  
TECHNOLOGIES

AKT | TECHNOLOGIES | PUBLICATIONS | RELATED PROJECTS | PEOPLE

ACQUISITION | MODELLING | RETRIEVAL | REUSE | PUBLISHING | MAINTENANCE

**3store from The University of Southampton**

**3Store:** MySQL based triple store, currently holding over 25 million RDF triples used by a range of Knowledgeable Services developed within the AKT project.

**3store fact-file**

Owner : The University of Southampton

Researchers (listed alphabetically) : Dr Nicholas Gibbins [[Browse, RDF](#)], Stephen Harris [[Browse, RDF](#)]

Description : <http://inanna.ecs.soton.ac.uk/3store/>

Demonstration : <http://www.hyphen.info/>

Builds on : Resource Description Framework, C, MySQL

Used by : Visualisations for the CS AKTive Portal, KnoZilla, CS AKTiveSpace, Eprep, ONTOCOPI

Addresses challenges : Knowledge Retrieval

**What's the Problem?**

Many Semantic Web applications require large quantities of RDF triples to perform their reasoning over.

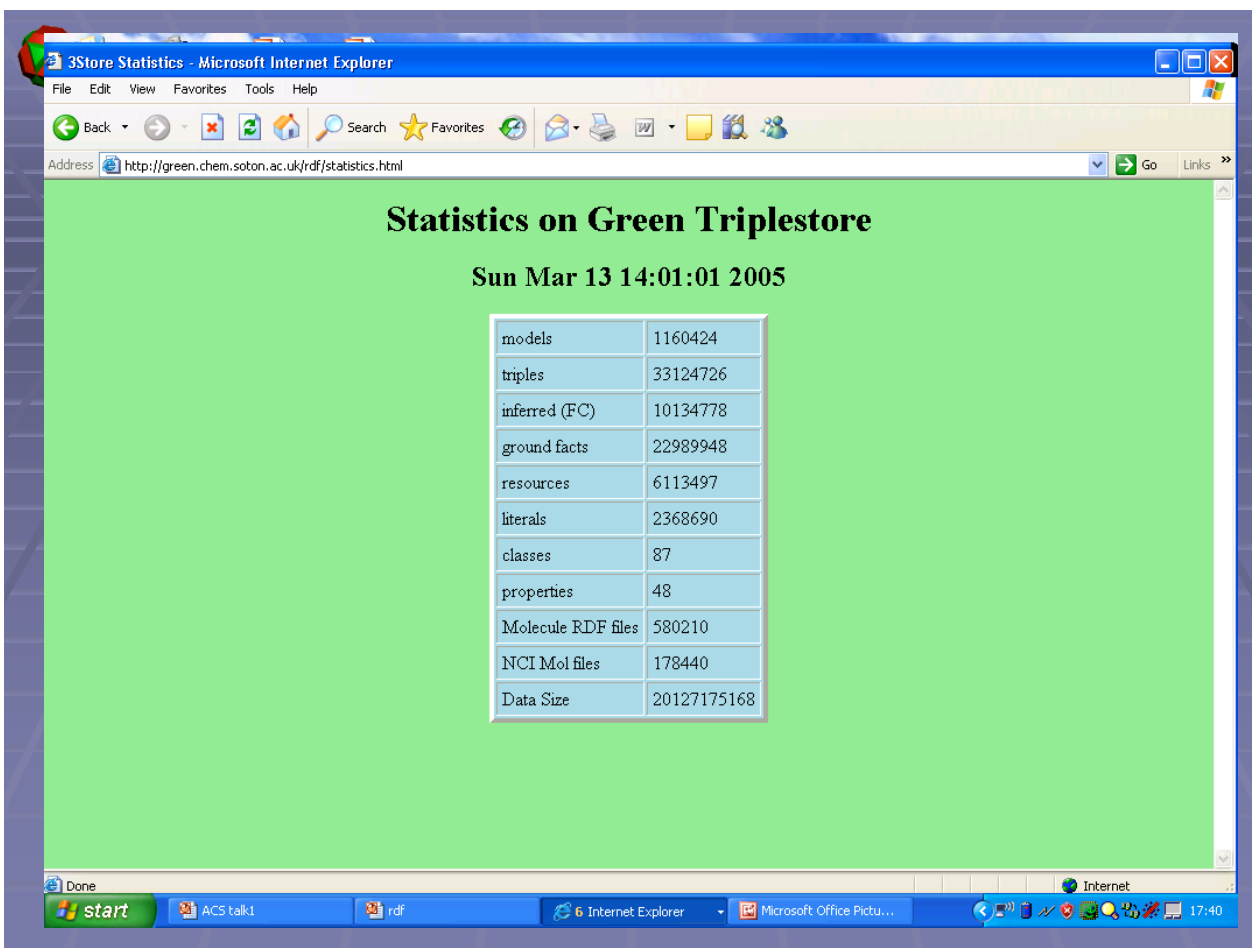
Current RDF database technologies scale to hundreds of thousands of triples.

start | Microsoft Office... | for JGF | My Computer | Webmail :: Login - ... | AKT - Technologies... | 07:05



## Triple Store – 3Store

- Need to have a triple store that scales
- 3Store from the AKT project looked like it would support this quantity of data
- Current tests suggest that it can.
- Achieves this by careful construction of a type of index
- Queries need to be done in RDQL
  - This is not yet fully developed but should be W3C standard soon



3Store Statistics - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://green.chem.soton.ac.uk/rdf/statistics.html> Go Links

## Statistics on Green Triplestore

Sun Mar 13 14:01:01 2005

models	1160424
triples	33124726
inferred (FC)	10134778
ground facts	22989948
resources	6113497
literals	2368690
classes	87
properties	48
Molecule RDF files	580210
NCI Mol files	178440
Data Size	20127175168

Done Internet

start ACS talk1 rdf Internet Explorer Microsoft Office Pictu... 17:40

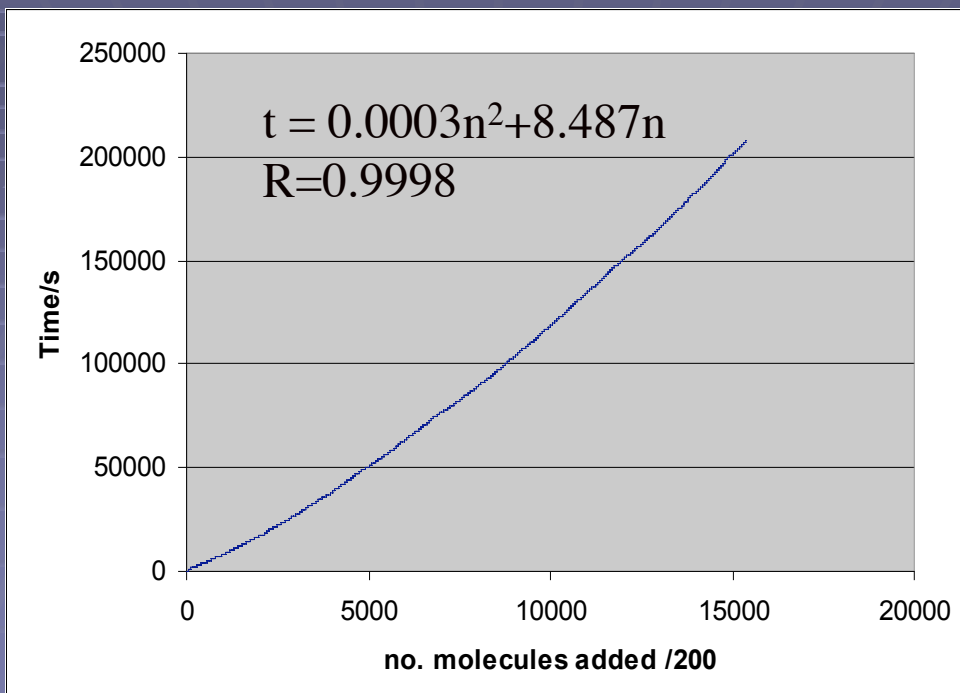


## Triple Stores

- Must be walked as the data structure is not determined in advance
  - Slower than relational database
  - Must interrogate rather than request



# Scaling?



March 2005

Jeremy Frey

ACS

**TRIANGLE**

File Edit View Go Bookmarks Tools Window Help

file:///rob/prog/triangle/test2/nitrotyrene-page.htm

File Edit View Go Bookmarks Tools Window Help

file:///rob/prog/triangle/test2/nitrotyrene-page.htm

CAS INCHI NAME SMILES EMPIR

1202-32-0

Search the DB Canned Example

query number = [10]

**MOLECULE**

type	Resource
type	Molecule
has-SMILES	c1ccc(cc1)C=C(C)N(=O)N=O
has name	B-BETHYL-B-NITROSTYR...
has-mpm	177.2
has-sterocentres	0
has-empirical-formula	C10H11NO2
has-cas	1202-32-0

**has-property**

- 48.29587
  - type PhysicalProperty
  - type BoilingPoint
  - type Resource
- has-information
  - 49.29587
    - type Resource
    - type Information
    - has-date 2005-2-28
    - has-value 307.5
  - has-uncertainty
  - has-unit
  - has-source
    - 53.29587
      - type Source
      - type PhysProp
      - type Resource
    - has-method
      - 54.29587
        - type Resource
        - type Laboratory
    - of-quality
    - has-condition
  - has-property
    - 16.29587
      - type PhysicalProperty
      - type Solubility
      - type Resource

**has-property**

  - 20.29587
    - type Resource
    - has-information
      - 6.29587
        - type Resource
        - has-information

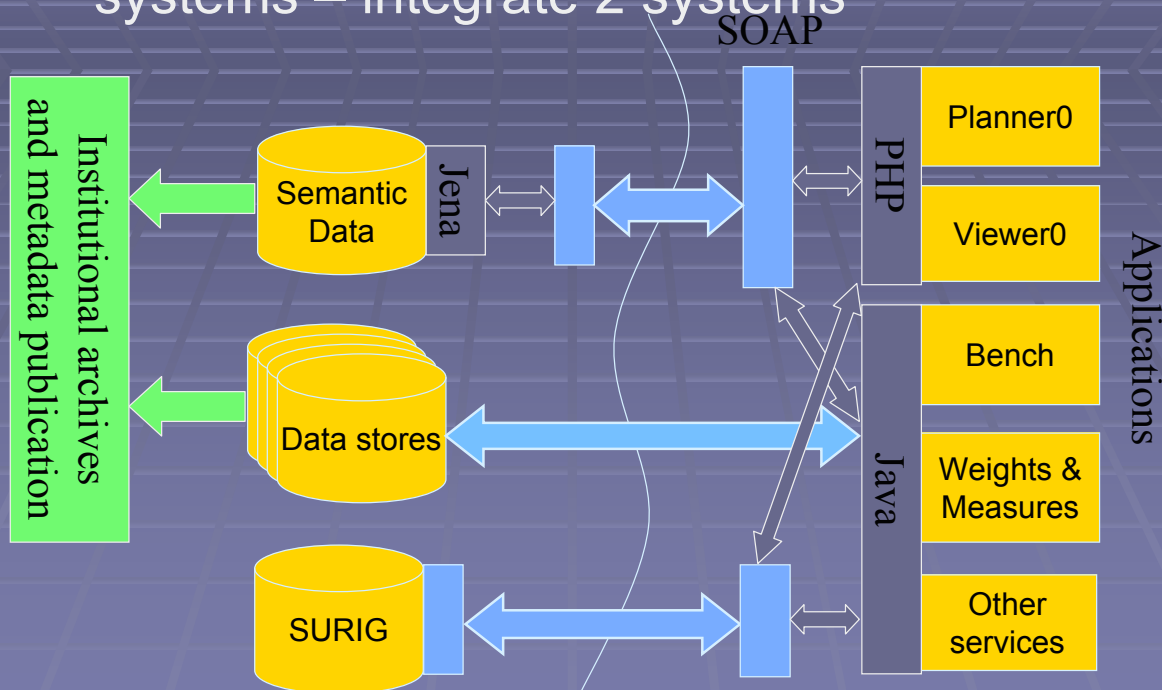
3D ball-and-stick model of the molecule.

© University of Southampton 2005  
ACCESS RESTRICTED  
Version 4.04

March 2005



# Similar architecture to the smart lab systems – integrate 2 systems



March 2005

Jeremy Frey

ACS



## The time is right to think of the Future – now we have an RDF Platform for experimental & computational combinatorial Chemistry

Now is the time to develop the asteroid deflection system, we don't need to wait for the full ontology

March 2005

Jeremy Frey

ACS



# People

- Kieron Taylor (Chemistry)
- Rob Gledhill (Chemistry)
- Hongchen Fu (Chemistry)
- Jamie Robinson (Chemistry)
- Steve Harris (ECS)
- Hugo Mills (ECS)
- Gareth Hughes (ECS)
- Jon Essex (Chemistry)
- Dave De Roure (ECS)
- Nigel Shadbolt (ECS)

**Making sure other people can re-use  
your data easily and with confidence**

Even when there is a huge amount of it!

But how big a triple store will we need?