Example Applications

Grid Computing (M)

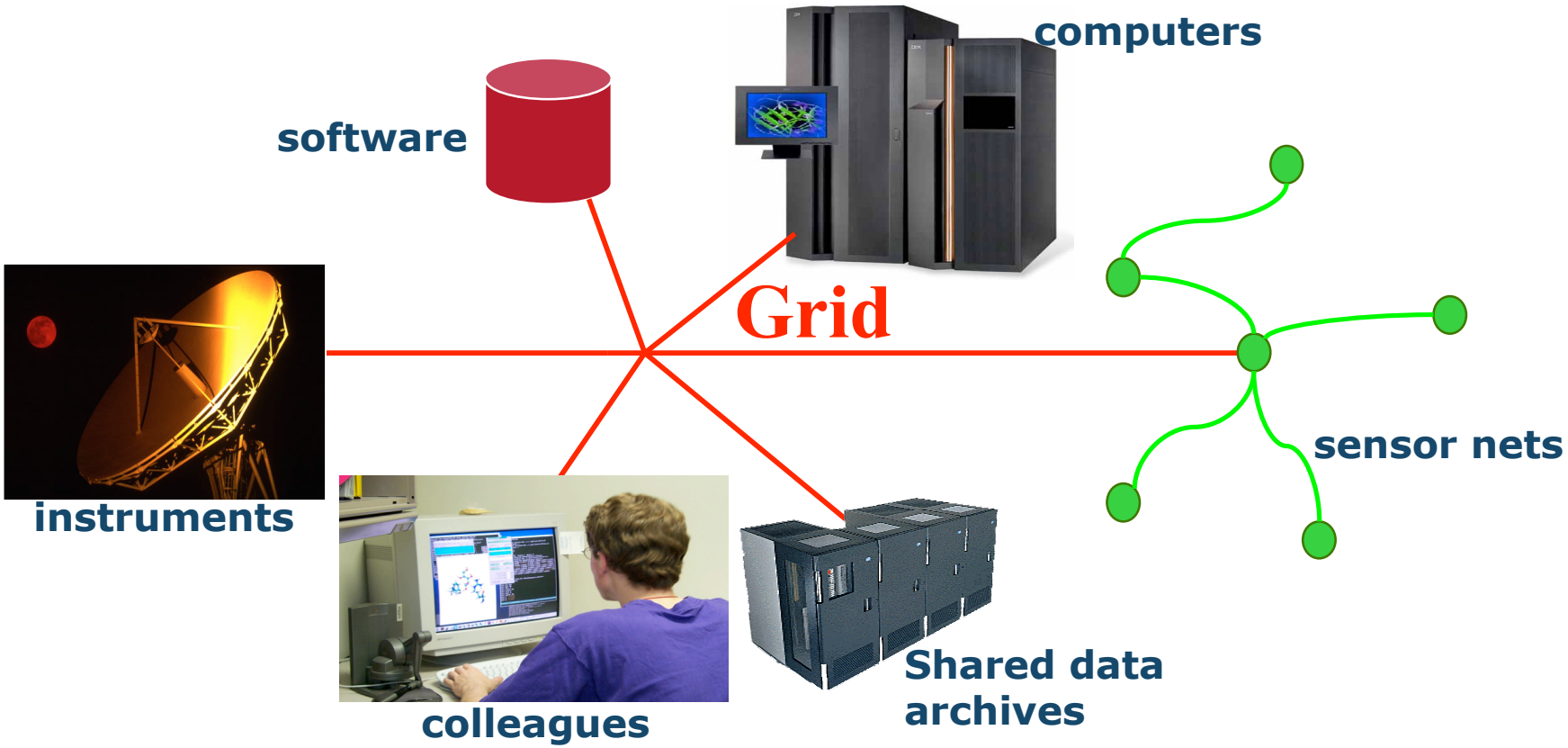
Prof. Richard Sinnott

UNIVERSITY
of
GLASGOW



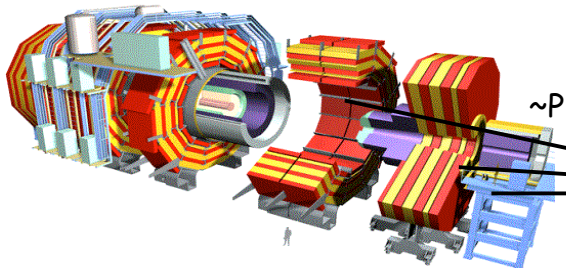
Foundation for e-Science

- e-Science methodologies transforming science, engineering, medicine and business
  - driven by exponential growth in data, compute demands
    - enabling a whole-system approach



Copyright © 2006 University of Glasgow

Data Grids for High Energy Physics



~PBytes/sec

Online System

~100 MBytes/sec

1 TIPS is approximately 25,000 SpecInt95 equivalents

Offline Processor Farm
~20 TIPS

~100 MBytes/sec

There is a "bunch crossing" every 25 nsecs.
There are 100 "triggers" per second
Each triggered event is ~1 MByte in size

Tier 0

CERN Computer Centre

HPSS

~622 Mbits/sec
or Air Freight (deprecated)

Tier 1

France Regional Centre

Germany Regional Centre

Italy Regional Centre

FermiLab ~4 TIPS

~622 Mbits/sec

Tier 2

Caltech
~1 TIPS

Tier2 Centre
~1 TIPS

~622 Mbits/sec

Institute
~0.25TIPS

Physics data cache

~1 MBytes/sec

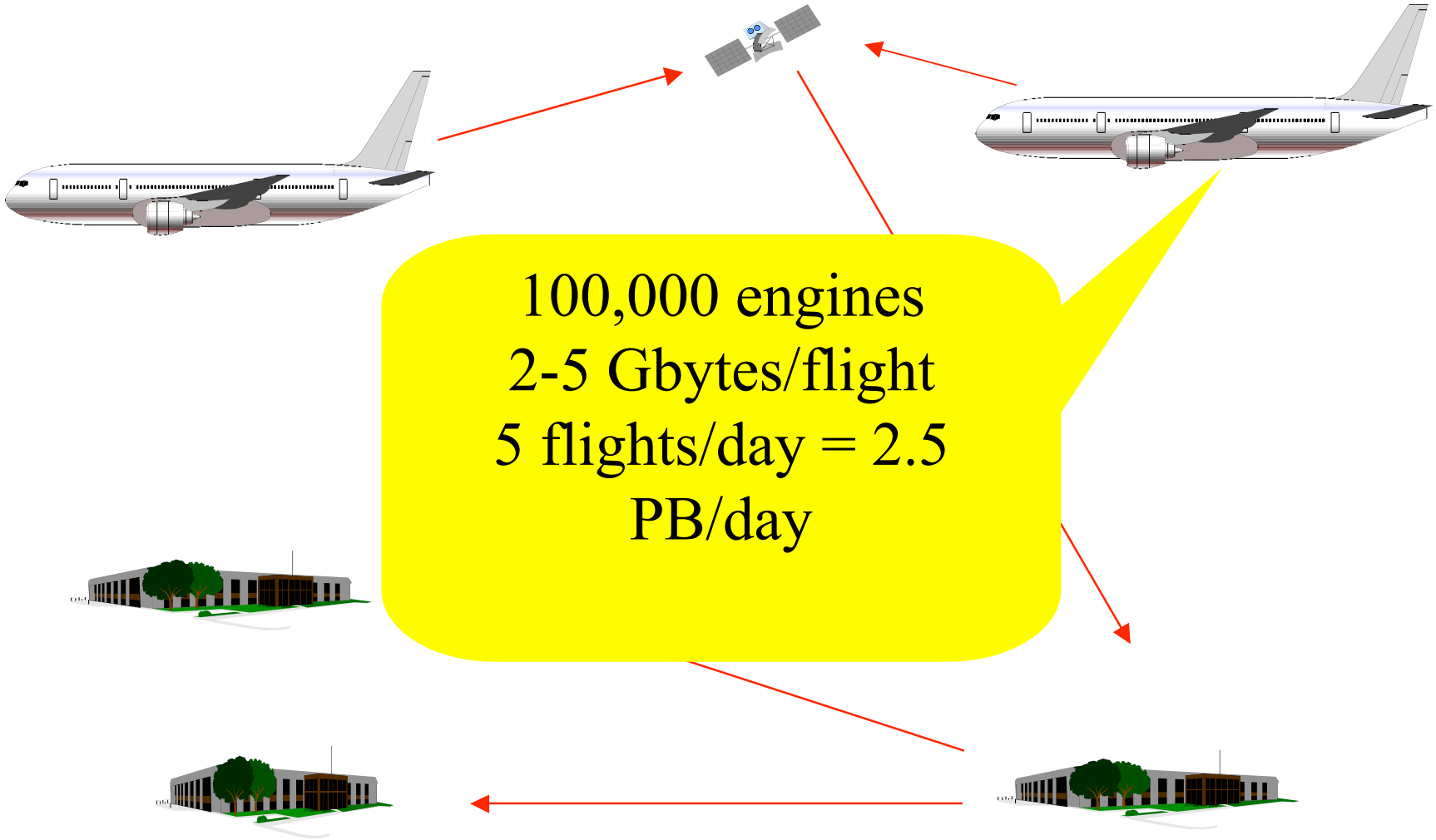
Tier 4

Physicist workstations

Physicists work on analysis "channels".
Each institute will have ~10 physicists working on one or more channels; data for these channels should be cached by the institute server

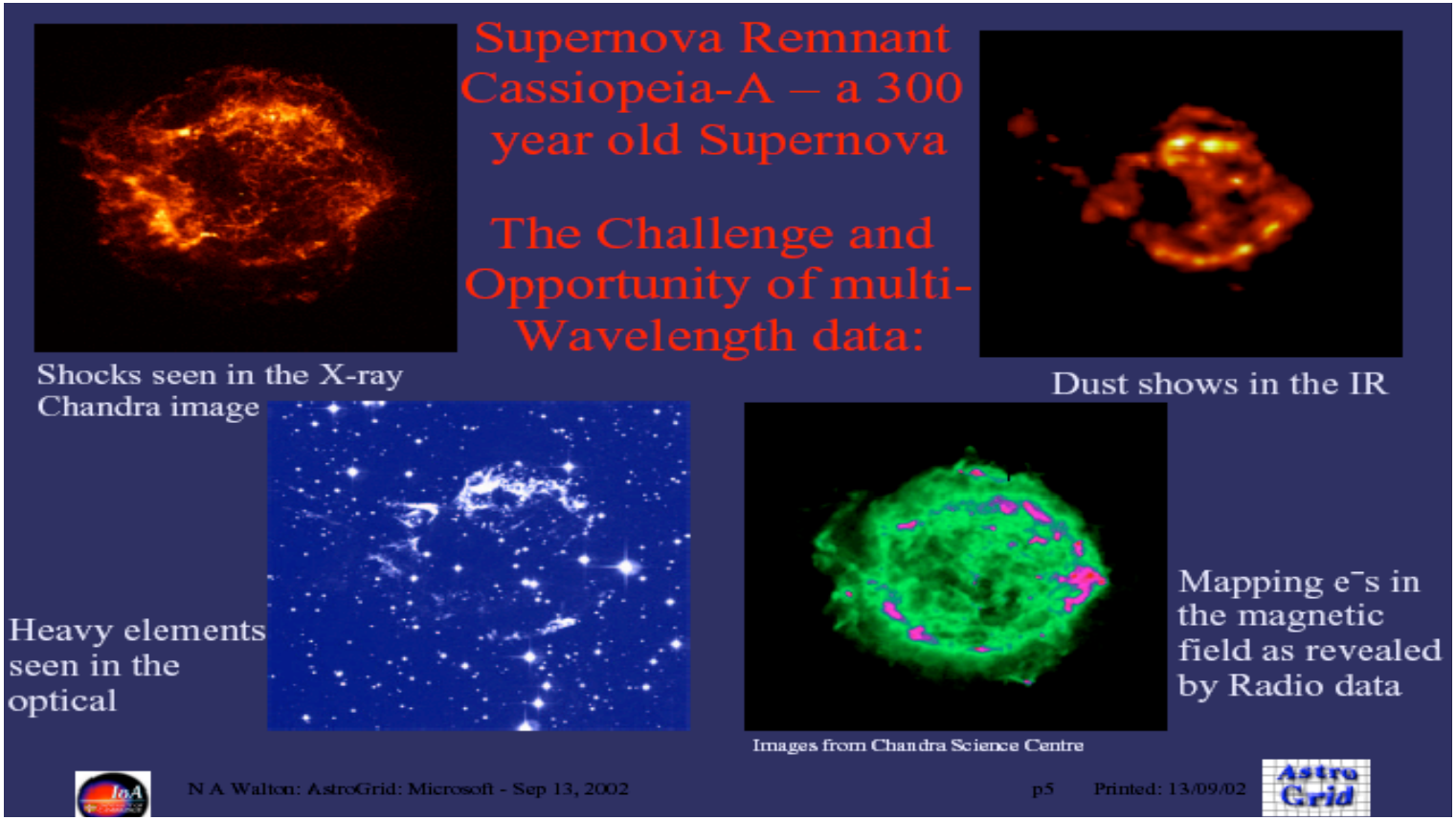
Copyright © 2006 University of Glasgow

Global in-flight engine diagnostics



Copyright © 2006 University of Glasgow

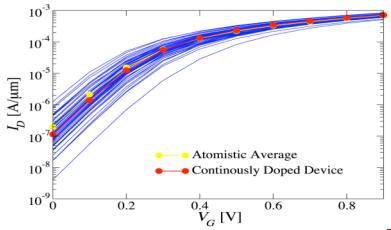
Virtual Observatories

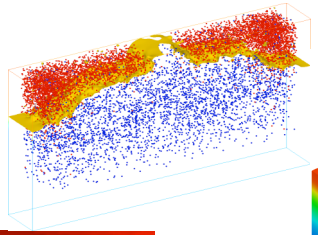


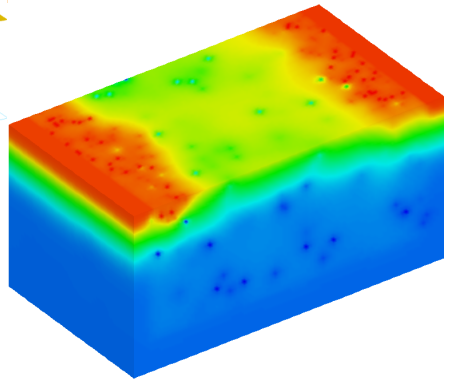
- Huge data sets
  - AstroGrid over 15TB data first week online
- Huge computations
  - Cross referencing data
  - Remove all junk from data sets
    - satellites, aeroplanes…

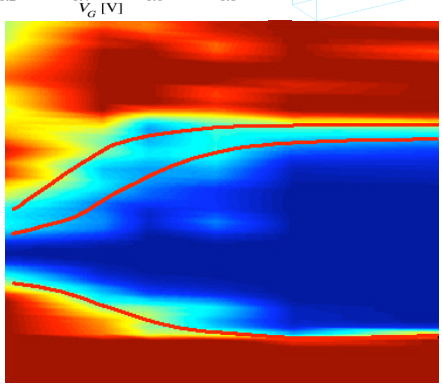
Copyright © 2006 University of Glasgow

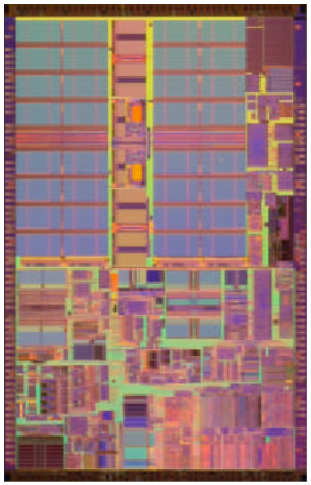
Next Generation Transistor Design







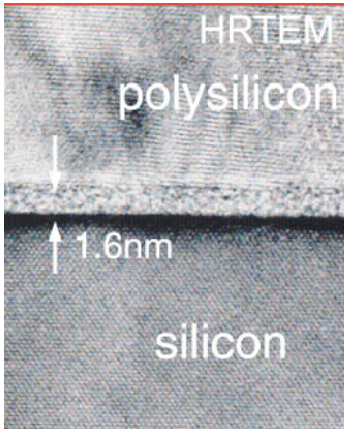


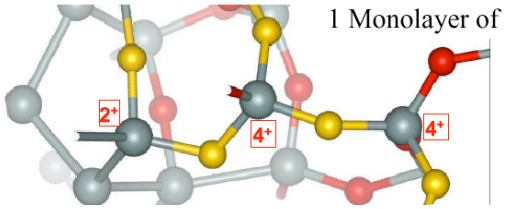


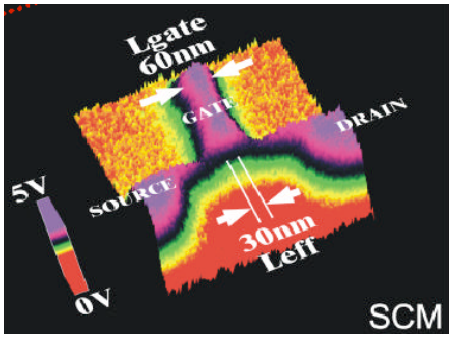
3D

+

Statistical







Copyright © 2006 University of Glasgow

Life Sciences

- Extensive Research Community
  - >1000 per research university
- Extensive Applications
  - Many people care about them
    - Health, Food, Environment
- Interacts with virtually every discipline
  - Physics, Chemistry, Maths/Stats, Nano-engineering, …
- 450+ databases relevant to bioinformatics (and growing!)
  - Heterogeneity, Interdependence, Complexity, Change, …
- Wonderful Scientific Questions
  - How does a cell work?
  - How does a brain work?
  - How does an organism develop?
  - What happens to the biosphere when the earth warms up?
  - Why do people who eat less tend to live longer?
  - …

Copyright © 2006 University of Glasgow



Yersinia pestis



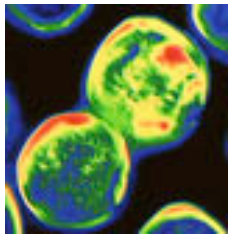
Arabidopsis thaliana



Buchnerasp. APS



Aquifex aeolicus



Archaeoglobus fulgidus



Borrelia burgorferi



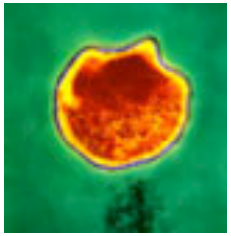
Mycobacterium tuberculosis



Caenorhabitis elegans



Campylobacter jejuni



Chlamydia pneumoniae



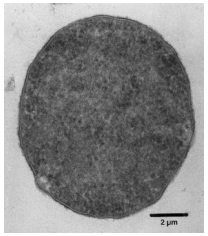
Vibrio cholerae



Drosophila melanogaster



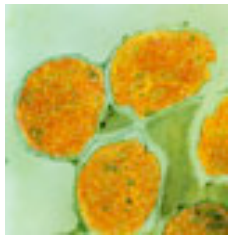
Escherichia coli



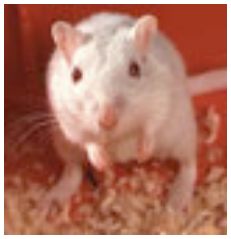
Thermoplasma acidophilum



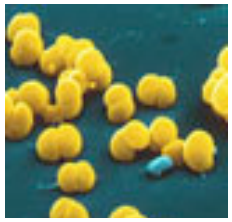
Helicobacter pylori



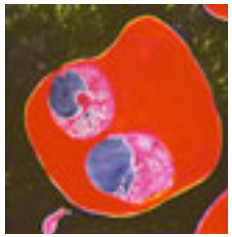
Mycobacterium leprae



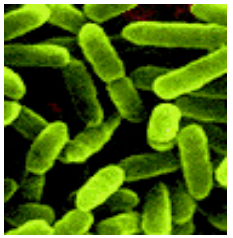
mouse



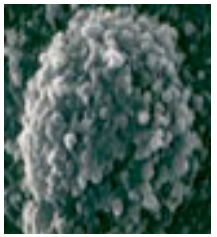
Neisseria meningitidis Z2491



Plasmodium falciparum



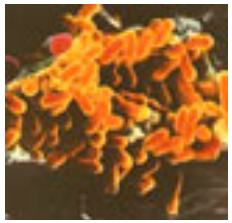
Pseudomonas aeruginosa



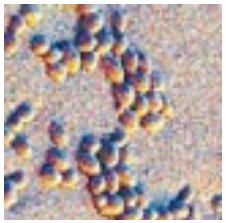
Ureaplasma urealyticum



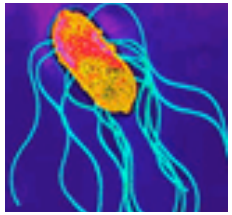
rat



Rickettsia prowazekii



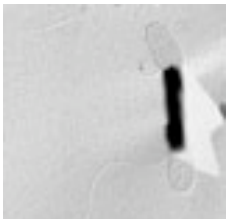
Saccharomyces cerevisiae



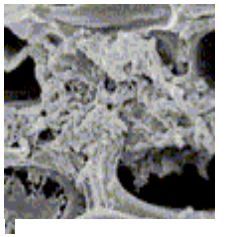
Salmonella enterica



Bacillus subtilis



Thermotoga maritima



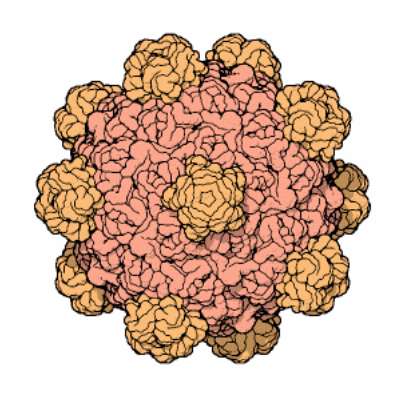
Xylella fastidiosa

Distributed and Heterogeneous data

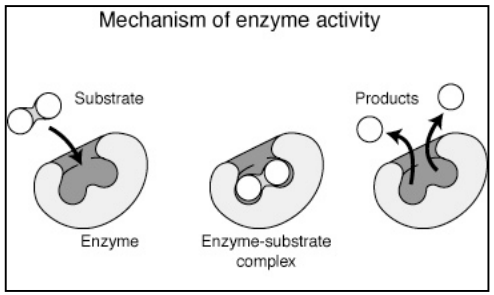
Sequence

```
LPSYVDWRSA GAVVDIKSQG
ECGGCWAFSA IATVEGINKI
TSGSLISLSE QELIDCGRTQ
NTRGCDGGYI TDGFQFIIND
GGINTEENYP YTAQDGDCDV
```

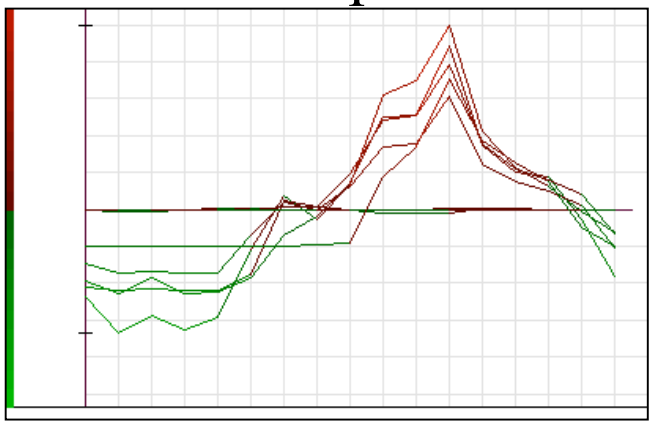
Structure



Function



Gene expression

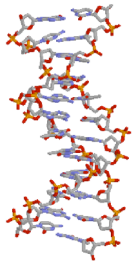


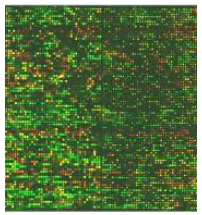
Morphology

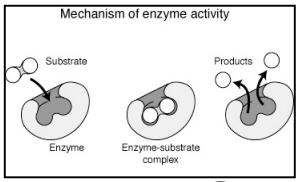


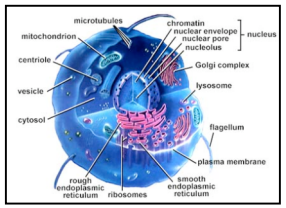
Copyright © 2006 University of Glasgow

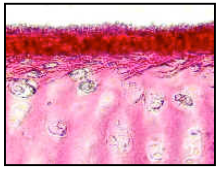
Systems-Biology…

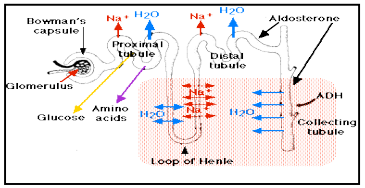












Nucleotide sequences

Nucleotide structures

Gene expressions

Protein Structures

Protein functions

Protein-protein interaction (pathways)

Cell

Cell signalling

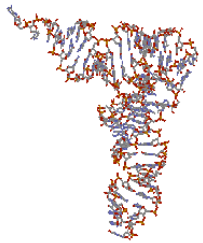
Tissues

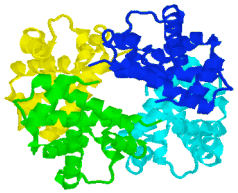
Organs

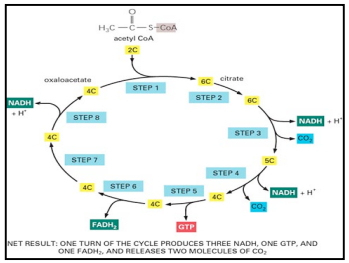
Physiology

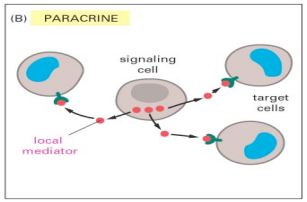
Organisms

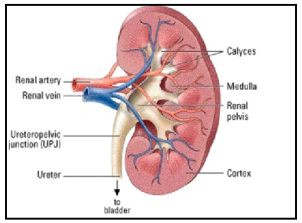
Populations

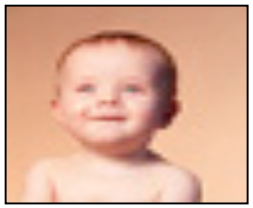










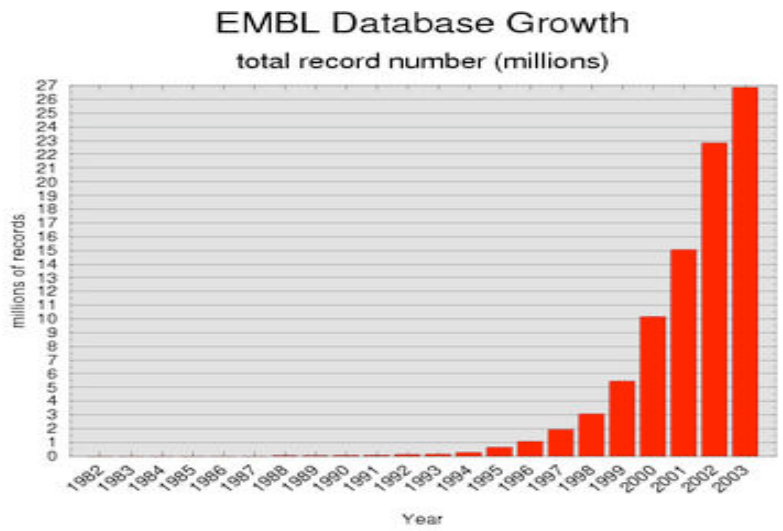


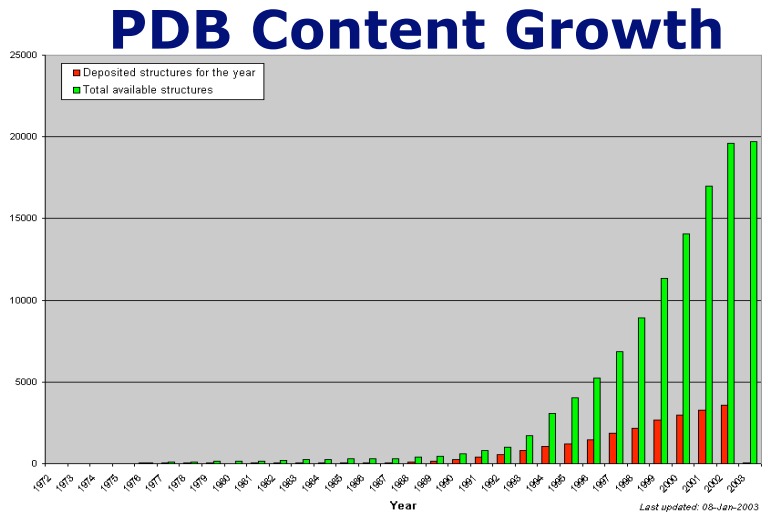


+ links to plant/crops, environmental, health, … information sources

Copyright © 2006 University of Glasgow

Database Growth





- DBs growing exponentially!!!
  - Biobliographic (MedLine, …)
  - Amino Acid Seq (SWISS-PROT, …)
  - 3D Molecular Structure (PDB, …)
  - Nucleotide Seq (GenBank, EMBL, …)
  - Biochemical Pathways (KEGG, WIT…)
  - Molecular Classifications (SCOP, CATH,…)
  - Motif Libraries (PROSITE, Blocks, …)

Copyright © 2006 University of Glasgow

Is Grid the Answer?

- Key problems to be addressed
  - Tools that *simplify* access to and usage of data
    - Internet hopping is not ideal!
  - Tools that *simplify* access to and usage of large scale HPC facilities
    - **qsub** [-a date_time] [-A account_string] [-c interval] [-C directive_prefix] [-e path] [-h] [-I] [-j join] [-k keep] [-l resource_list] [-m mail_options] [-M user_list] [-N name] [-o path] [-p priority] [-q destination] [-r c] [-S path_list] [-u user_list] [-v variable_list] [-V] [-W additional_attributes] [-z] [script]
  - Tools designed to *aid understanding* of complex data sets and relationships between them
    - e.g. through visualisation

Copyright © 2006 University of Glasgow

Access to and Usage of Data

- Grid technology should allow to
  - hide heterogeneity,
  - deal with location transparency,
  - address security concerns,
  - …

- Data Access and Integration Specification (DAIS) being defined by GGF
  - OGSA-DAI and DAIT projects key role in shaping these standards
- Other commercial solutions (IBM Information Integrator, …)
  - More later!

Copyright © 2006 University of Glasgow

Access to and Usage of HPC facilities

- Consider whole genome-genome (2*3*10^9 bp) comparisons between two species
  - Current strategy essentially chops up one genome and fires searches for those fragments in the other then re-assembles results
    - messy approximate matching - re-assembly difficult
    - important correlations can be lost
      - to make this tractable so called junk DNA ignored
      - chopping may introduce artefacts or hide phenomena

➢**Better to put both full genomes in memory and perform a useful complete comparison**
➢**Only possible with very high-end machines** (available via grids)

  - Should not have to be script writer/Linux sys-admin to use these facilities

Copyright © 2006 University of Glasgow

Cognitive aspects of Data

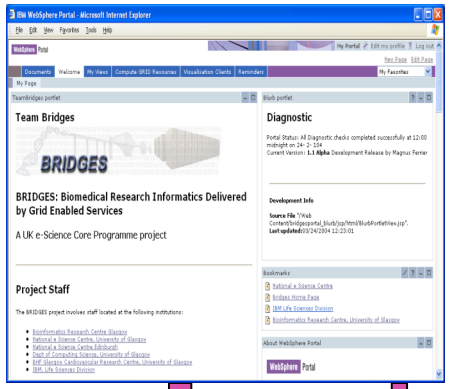
- Life science data can be “ugly”
  - Raw data sets messy
  - Requires significant effort to understand
  - Schemas/data models evolving
  - …

- Tools needed to
  - Simplify understanding
  - Improve analysis
  - Navigate through potentially huge data sets
    - e.g. to find genes of interest in chromosomes of different species

Copyright © 2006 University of Glasgow

BRIDGES Project

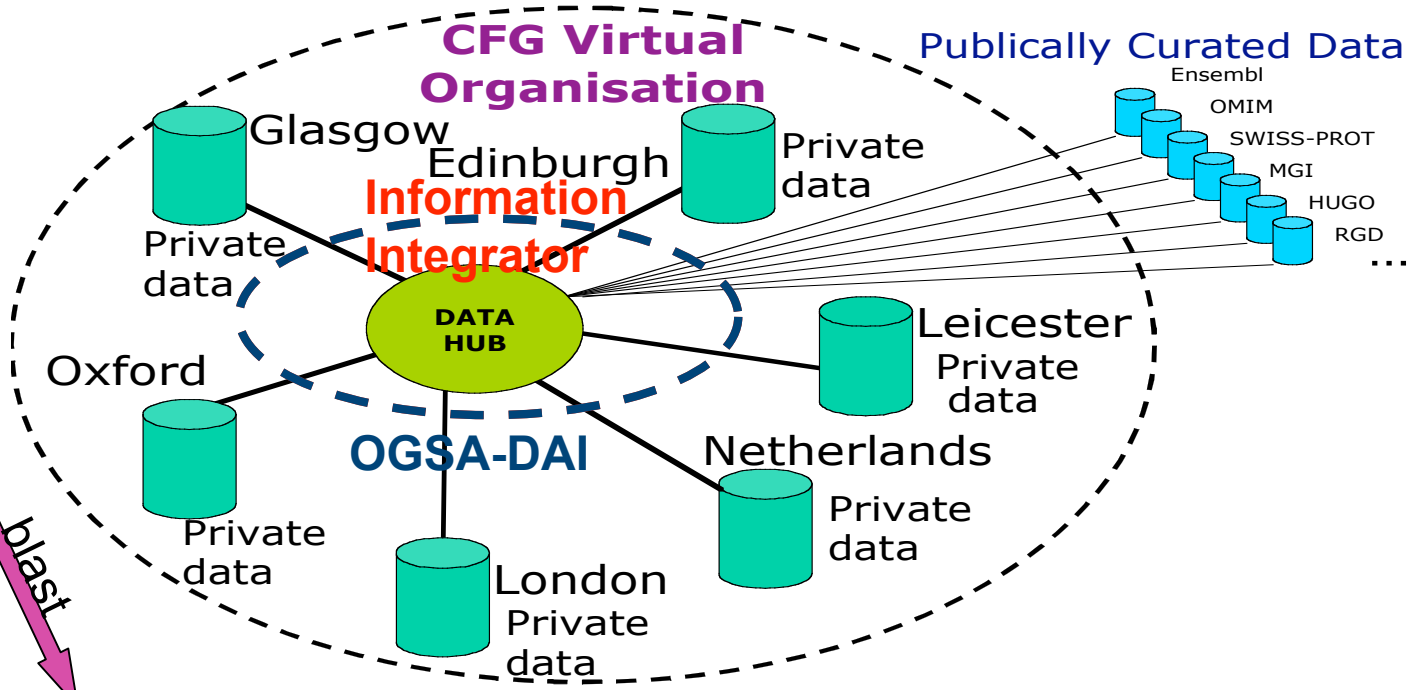
VO Authorisation

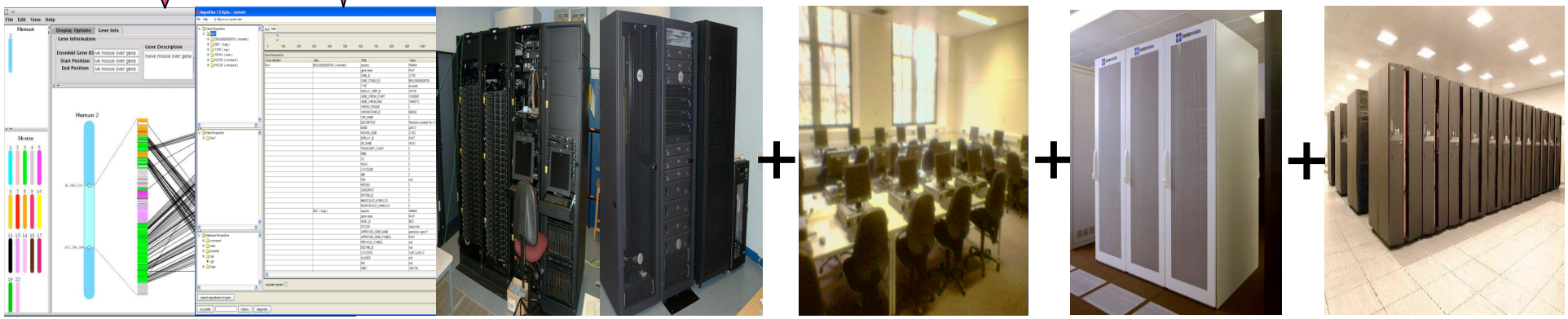


Synteny Service

Magna Vista Service

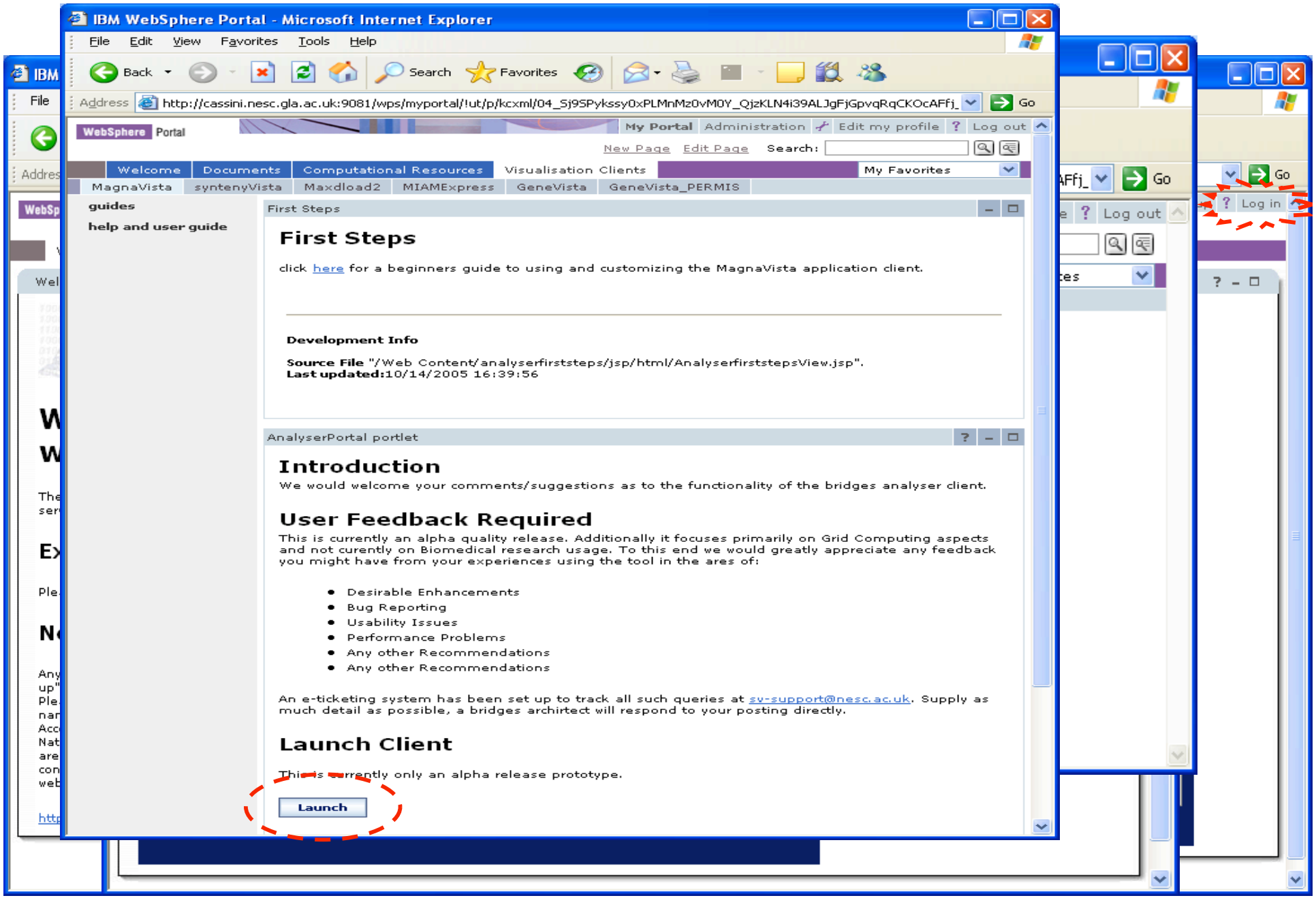
blast



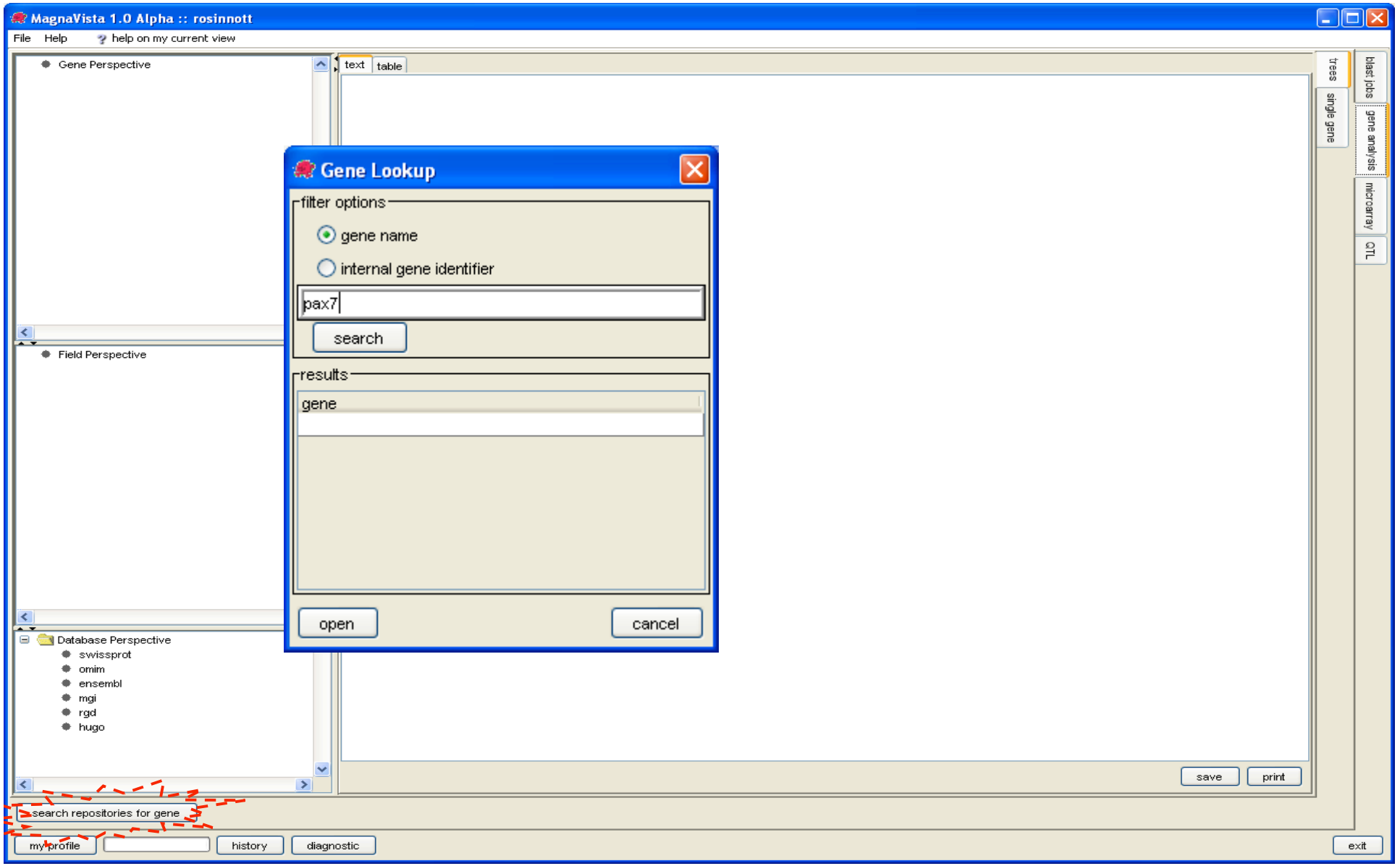


Copyright © 2006 Uni

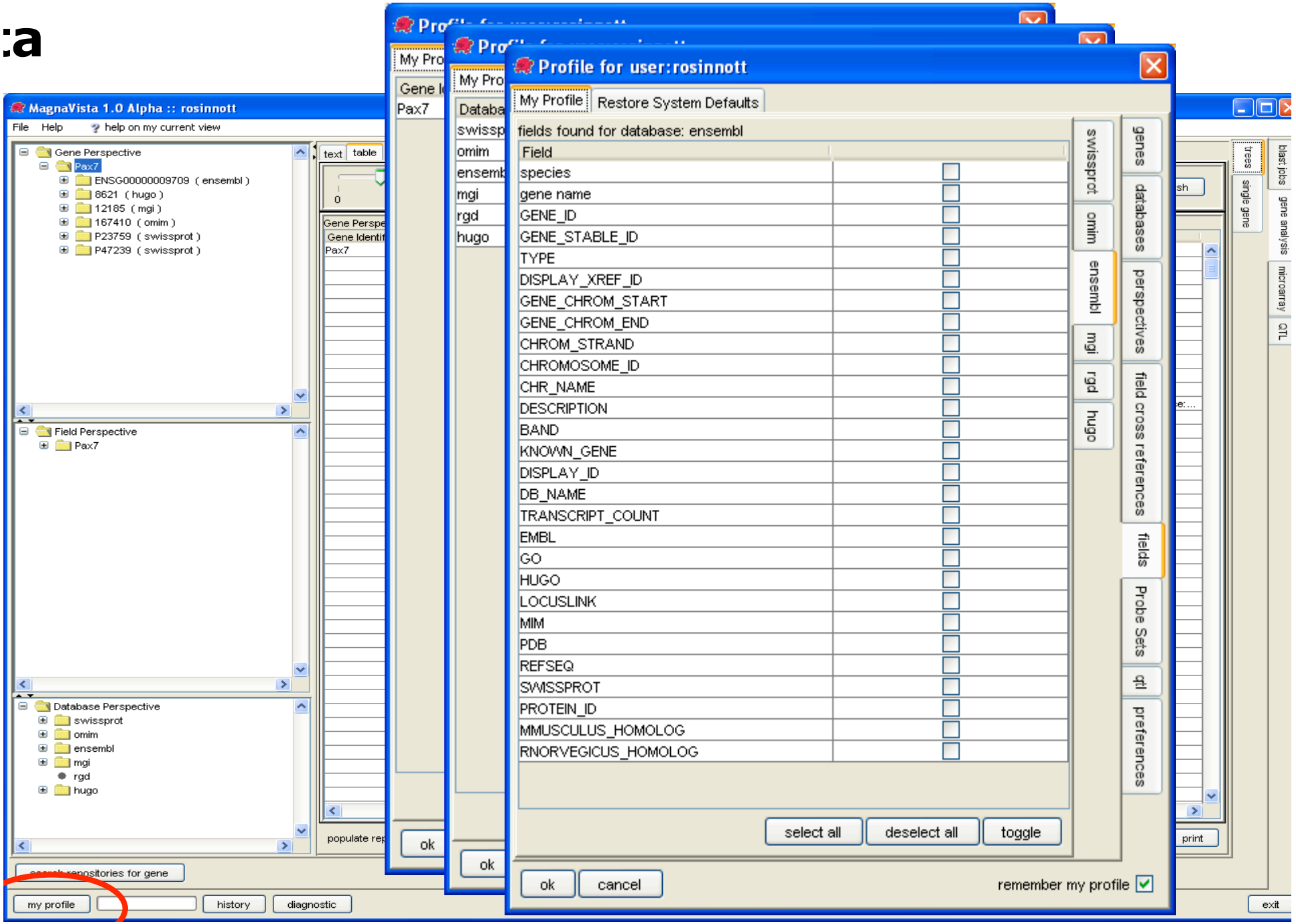
Bridges Portal



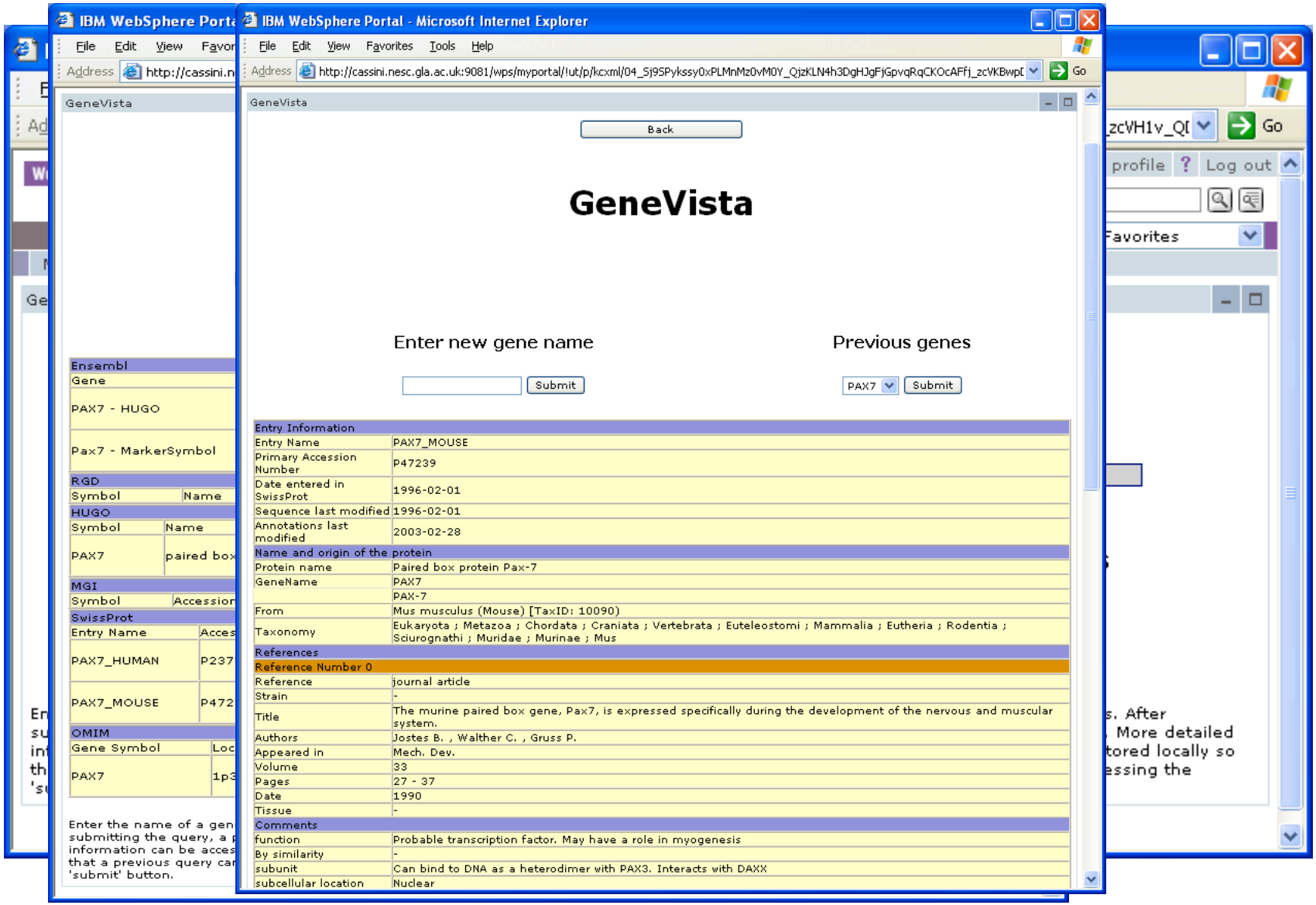
MagnaVista


MagnaVista 1.0 Alpha :: rosinnott
File
Help
help on my current view
Gene Perspective
Field Perspective
Database Perspective
swissprot
omim
ensembl
mgi
rgd
hugo
text
table
trees
single gene
blast jobs
gene analysis
microarray
QTL
Gene Lookup
filter options
gene name
internal gene identifier
pax7
search
results
gene
open
cancel
save
print
search repositories for gene
my profile
history
diagnostic
exit

:a

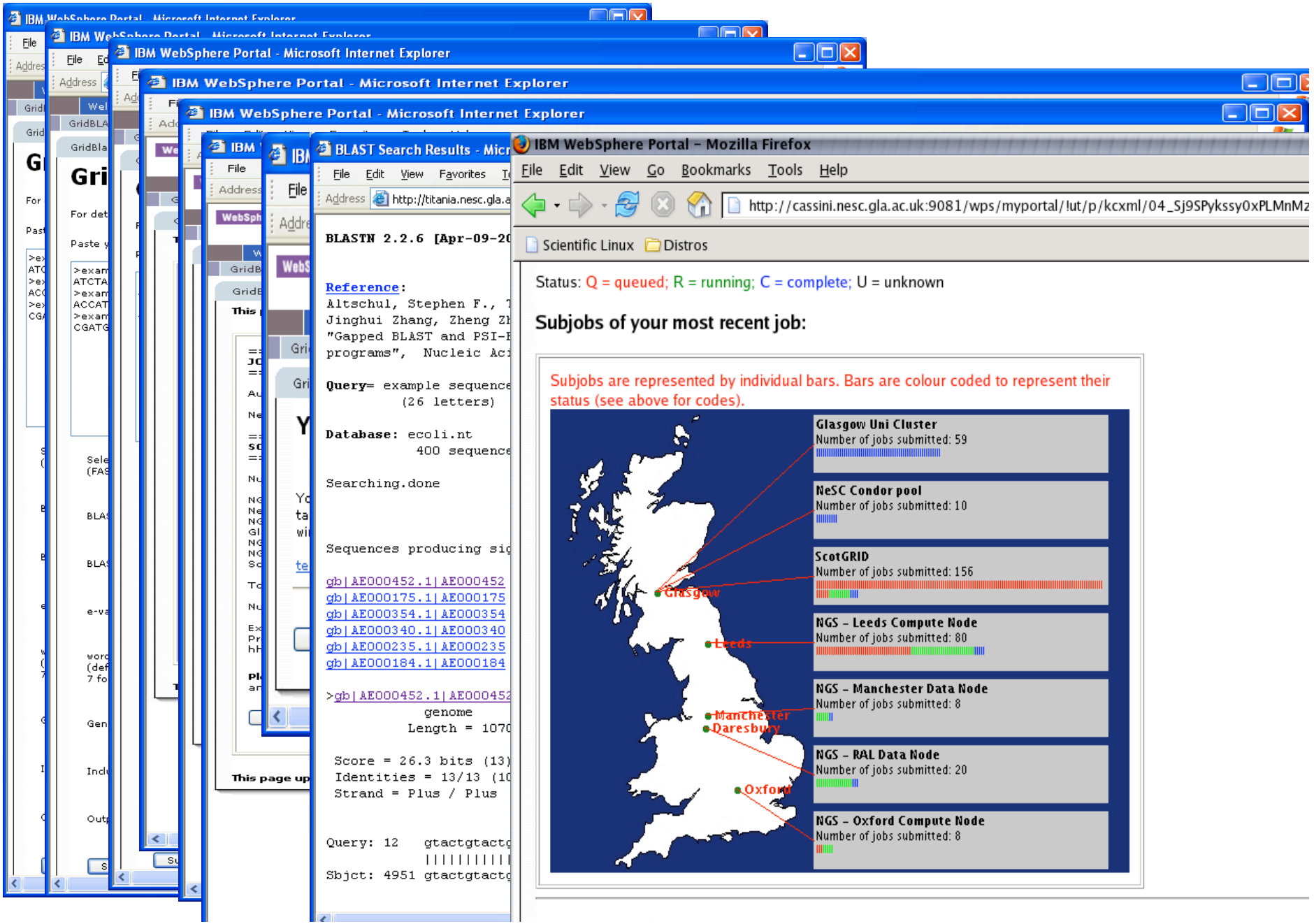

MagnaVista 1.0 Alpha :: rosinnott
File
Help
help on my current view
Gene Perspective
Pax7
ENSG00000009709 (ensembl)
8621 (hugo)
12185 (mgi)
167410 (omim)
P23759 (swissprot)
P47239 (swissprot)
Field Perspective
Pax7
Database Perspective
swissprot
omim
ensembl
mgi
rgd
hugo
text
table
Profile for user:rosinnott
My Profile
Restore System Defaults
fields found for database: ensembl
Field
species
gene name
GENE_ID
GENE_STABLE_ID
TYPE
DISPLAY_XREF_ID
GENE_CHROM_START
GENE_CHROM_END
CHROM_STRAND
CHROMOSOME_ID
CHR_NAME
DESCRIPTION
BAND
KNOWN_GENE
DISPLAY_ID
DB_NAME
TRANSCRIPT_COUNT
EMBL
GO
HUGO
LOCUSLINK
MIM
PDB
REFSEQ
SWISSPROT
PROTEIN_ID
MMUSCULUS_HOMOLOG
RNORVEGICUS_HOMOLOG
select all
deselect all
toggle
swissprot
omim
ensembl
mgi
rgd
hugo
genes
databases
perspectives
field cross references
fields
Probe Sets
qtl
preferences
ok
cancel
remember my profile
my profile
history
diagnostic
exit
print
trees
single gene
blast jobs
gene analysis
microarray
QTL

GeneVista



Copyright © 2006 University of Glasgow

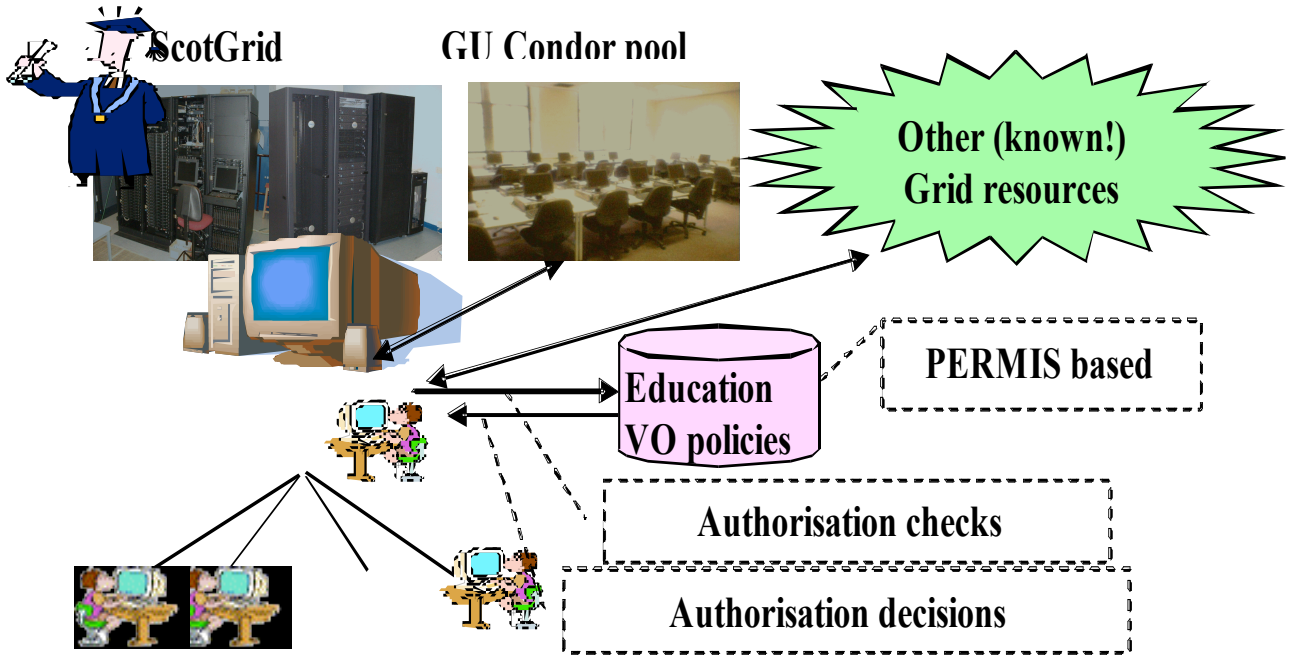
Grid Blast Interface



Copyright © 2006 University of Glasgow

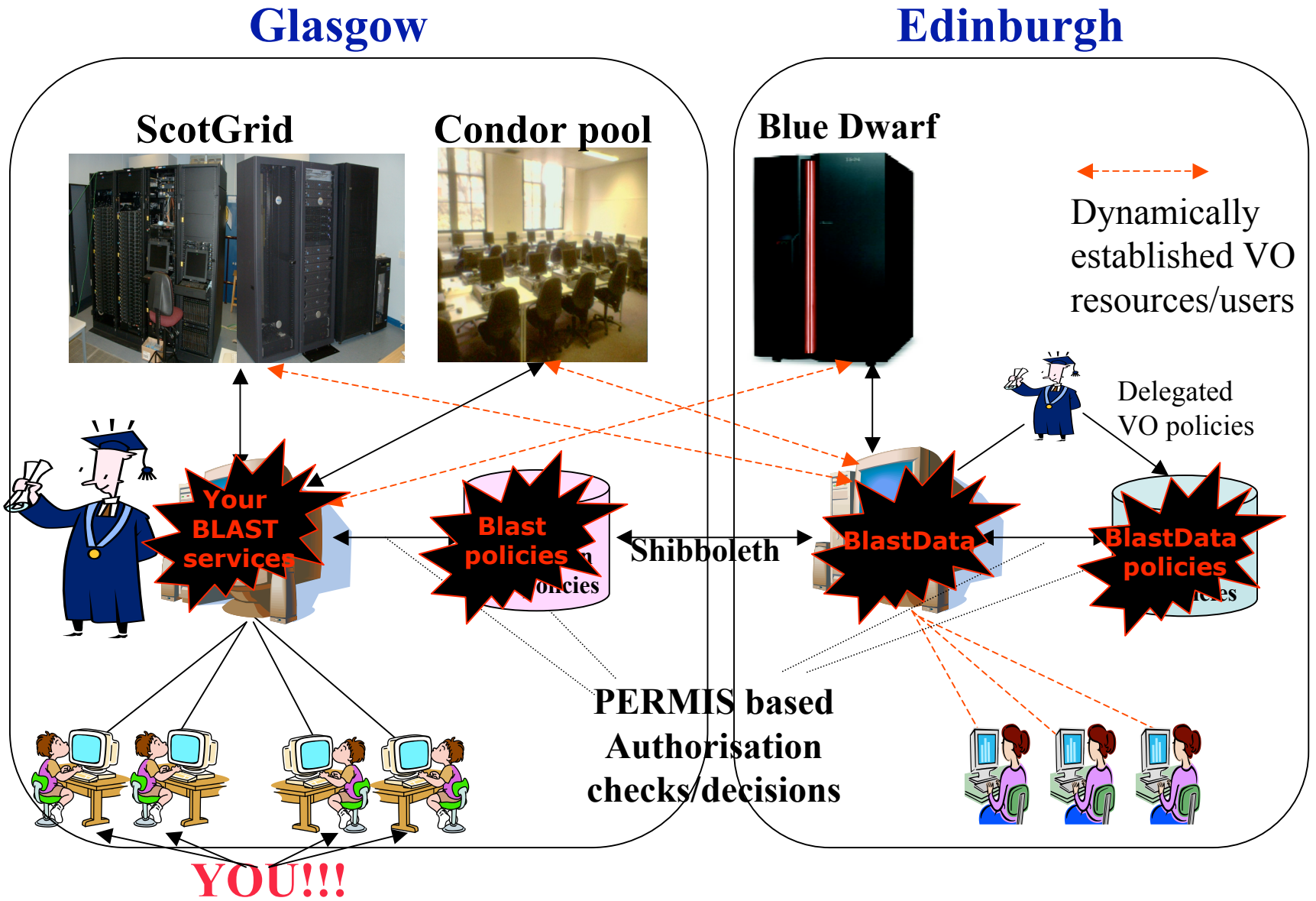
DyVOSE Project

- Dynamic Virtual Organisations for e-Science Education (DyVOSE) project
  - Two year project (£289k) started 1st May 2004 funded by JISC
  - Exploring advanced authorisation infrastructures for security
    - … in Grid Computing Module as part of advanced MSc at Glasgow
      - providing insight into rolling Grid out to the masses!



Copyright © 2006 University of Glasgow

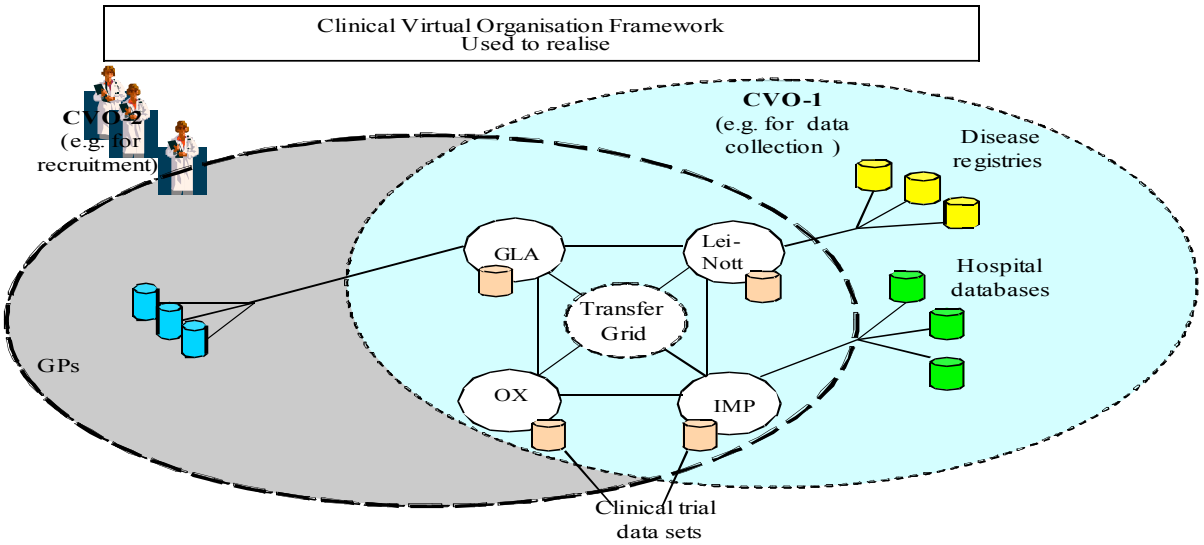
DyVOSE Phase 2/3



Copyright © 2006 University of Glasgow

VOTES

- Virtual Organisations for Trials and Epidemiological Studies
  - 3 year MRC (£2.8M) funded project started October 2005
  - Plans to develop Grid infrastructure to address key components of clinical trial/observational study
    - Recruitment of potentially eligible participants
    - Data collection during the study
    - Study administration and coordination
      - Involves Glasgow, Oxford, Leicester, Nottingham, Manchester



Copyright © 2006 University of Glasgow

Clinical Trials 101

- Need to answer questions such as
  - How many men in Scotland between the ages of 45-65 had a heart attack in the last 5 years? Of those that did, would they be interested in trialling a new drug to prevent possible further serious major events?
    - Recruitment!
  - For recruited men, are they regularly taking the new drug (or placebo)? Do they visit their GP/hospital regularly for the drug/placebo, to give samples, for monitoring purposes? Did they have any further major events (or side-effects) in taking the drug?
    - Data collection!
  - Who can see the information associated with this trial? Can a hospital doctor, nurse see all of given patients data? Only their GP? A clinical trials researcher? Who ensures that a study is in the patients interest? Can we simplify the ethical review process? Who checks the validity of trial results?
    - Study management!
  - Through this we would eventually like to be able to answer questions like,
    - is the high rate of heart attacks in Scotland caused by genetic factors
      - …or being an e-Science director…?

Copyright © 2006 University of Glasgow

Grid Possibilities…

- (Digital) data is out there in NHS land, other places …
  - More later…
- Grid technology provides glue to bring it together, but …
  - Need to ensure
    - Ethical and legal factors considered
      - Patient consent for the data to be made available
    - Right people seeing right data, in right context for right purpose, …
      - In moving data from site A to site C can it go via site B?
    - Right services made available
      - For finding, accessing, querying, analysing data…
        - » Note general publish model mostly not possible
    - Provenance of data
      - Where did it come from, how was it generated, has it been validated, …
- MRC funded VOTES project is exploring this space…

Copyright © 2006 University of Glasgow

Initial Experiences

- Focus initially on proof of concept prototypes
  - Easier to influence policy when working systems to show consent scenarios, why CHI needed, how Grid security actually works, … etc

- Roles
  - We've made some up to begin with…
    - When is a doctor not a doctor?
      - When he's a consultant, a registrar, a surgeon, a professor, a PhD, a…
        - » …you get the picture!
    - Is a Glasgow Staff Nurse Grade G = a Tayside Staff Nurse G
      - Is there an A-Z/ontology/dictionary/standardised description of these roles and the associated job descriptions
        - » ideally with various roles involved in clinical trials/observational studies?

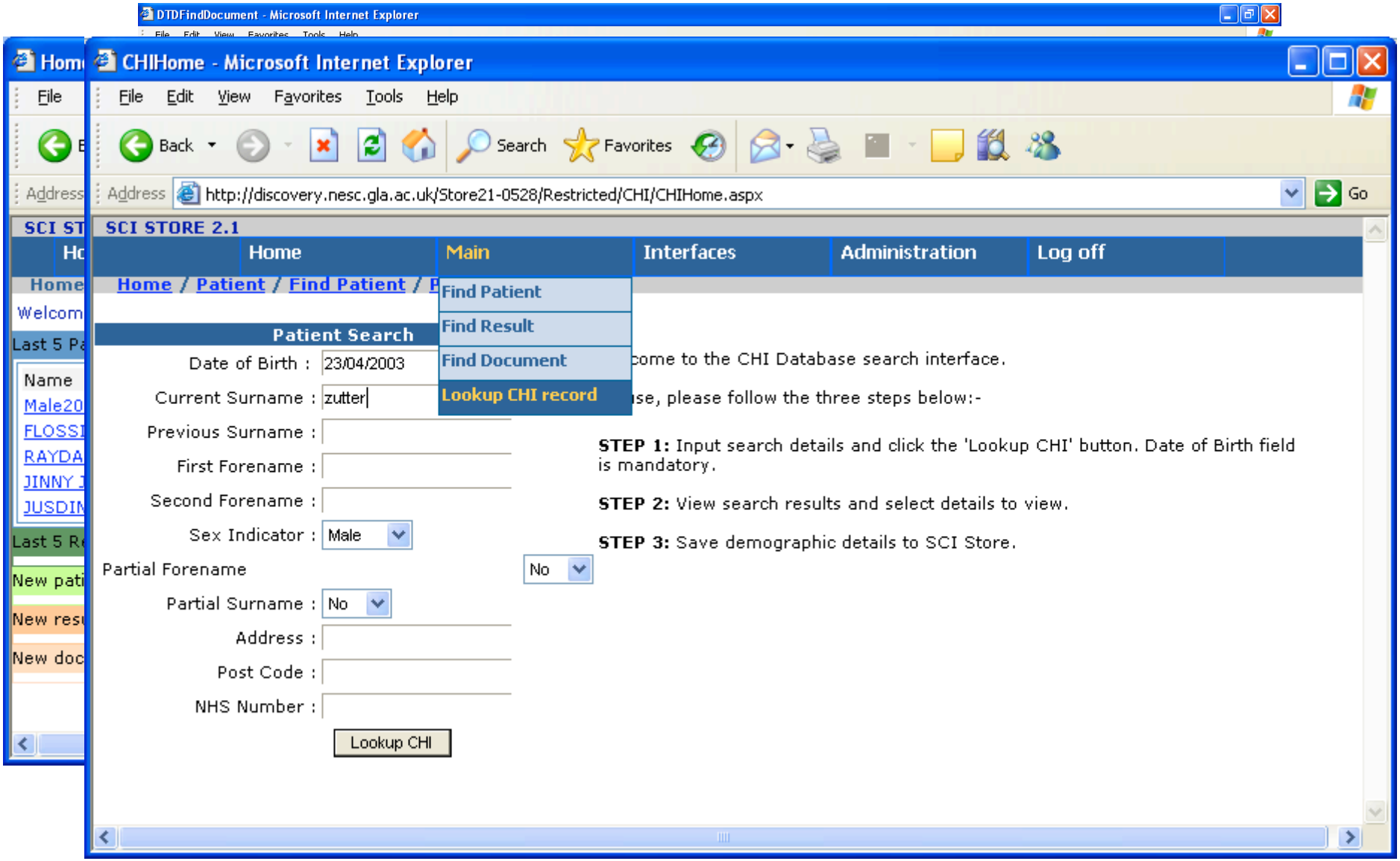
Copyright © 2006 University of Glasgow

Initial Experiences

- Scottish Data Space… up to now
  - Scottish Care Information (SCI) Store
  - Scottish Morbidity Records (SMR)
  - General Practitioners Administration System for Scotland (GPASS)
  - Data dictionary
  - …
  - Consent database

Copyright © 2006 University of Glasgow

A Quick Tour of SCI Store



Copyright © 2006 University of Glasgow

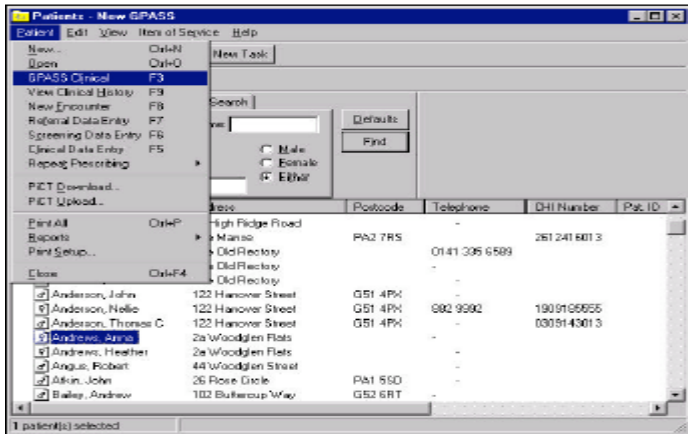
Scottish Morbidity Records

- Scottish Morbidity Records
  - Good quality data sets put together by ISD
    - Historic SMR1 Discharges January 1981 - March 1997
    - COPPISH SMR01 Discharges April 1997 onwards
    - Historic SMR4 Discharges 1981 – March 1997
    - COPPISH SMR04 Admissions April 1996 onwards
    - GRO Death Records January 1980 - December1995
    - GRO Death Records January 1996 onwards
    - SOCRATES (Cancer Registrations) 1980 onwards
  - Negotiating access to anonymised SMR data sets

Copyright © 2006 University of Glasgow

GPASS

- General Practice Administration System for Scotland (GPASS)
  - used by over 85% of GPs in Scotland
    - links from SCI Store to GPASS
  - access to GPASS software with training data sets
    - XML API available for querying
      - www.gpass.co.uk



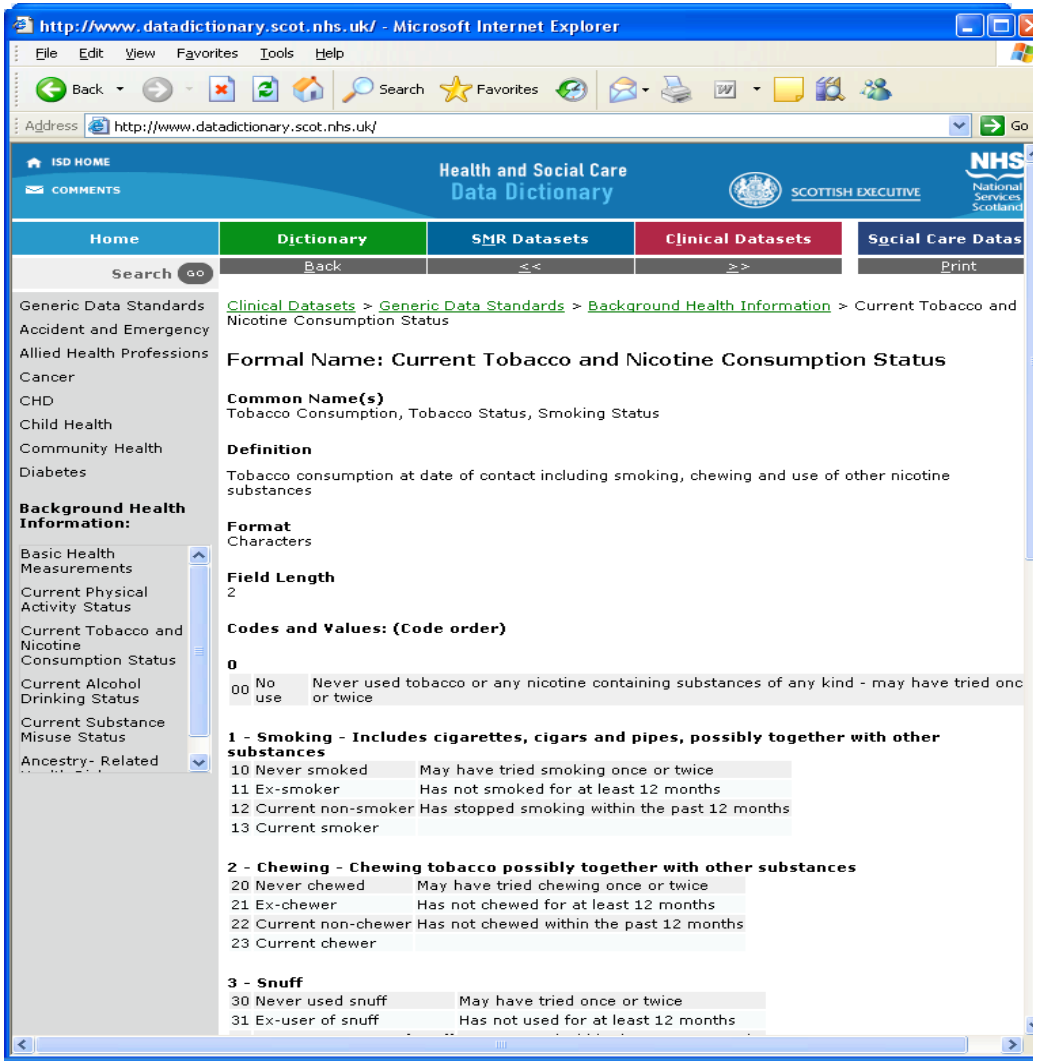
Copyright © 2006 University of Glasgow

Data Dictionary

- Includes vocabulary for
  - SMR data
  - Clinical data
  - Social care data

    - Negotiating access to DB back end or web service front end to this

    - Will link to data federation framework / tools



Copyright © 2006 University of Glasgow

Data Consent

- Simple solution implemented at NeSC for now
  - E-Health strategy states it will eventually be included as part of patient electronic health record in Scotland
    - When…?
- Scenarios based upon patient consent
  - For general use of data
  - For usage of data in particular trial
  - For usage of data in particular disease research
    - Combinations of these possible/likely!
  - Opt in vs Opt out?
    - We can implement various scenarios to show how it might work, but need to be aligned with "real world"
      - i.e. with NHS plans/strategy for e-Health

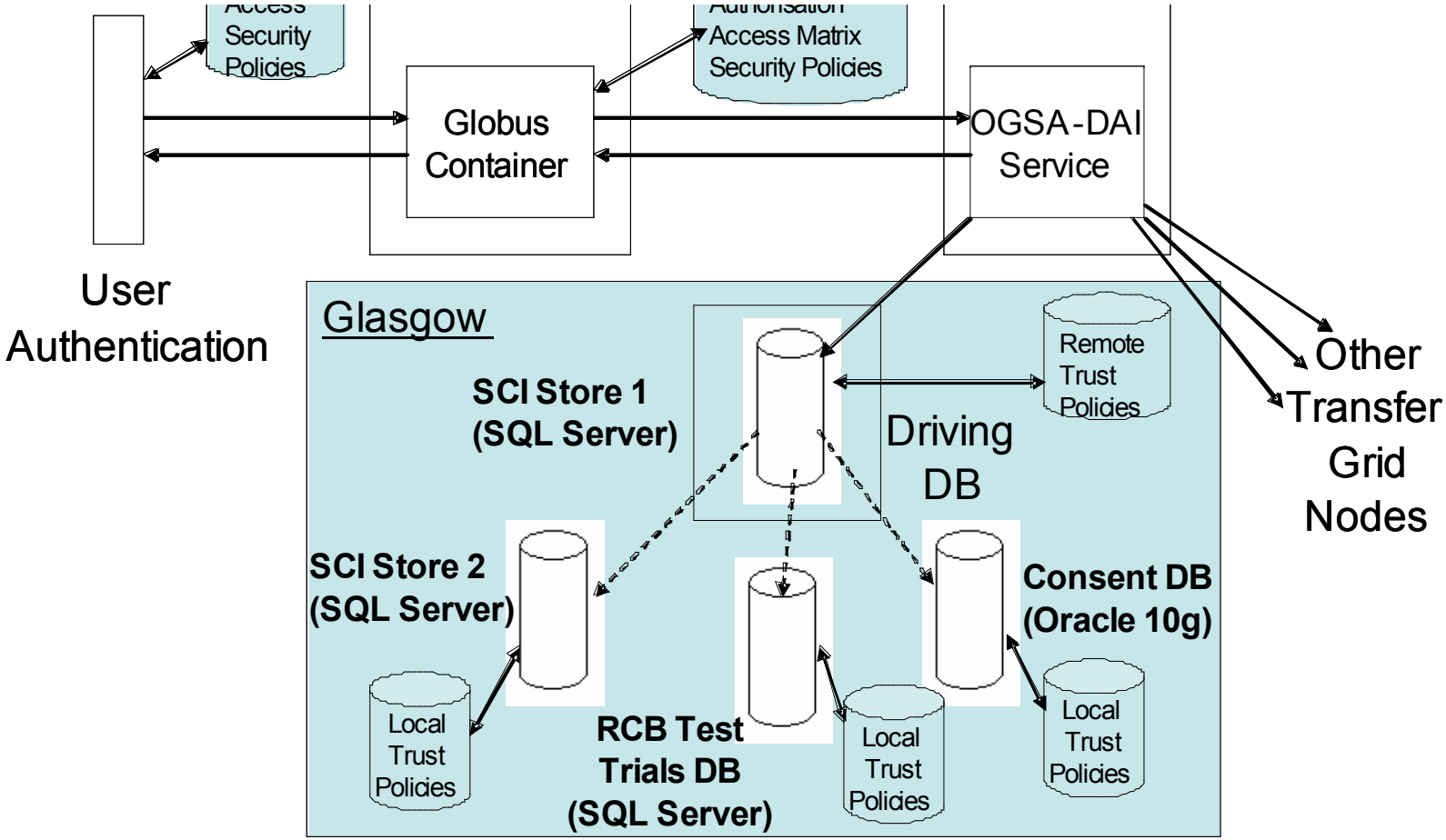
Copyright © 2006 University of Glasgow

Bringing it All Together…

- Technologies
  - Globus Toolkit (v4.0) for Grid services
  - OGSA-DAI for data federation
  - PERMIS for security
    - Access matrix (for rapid prototyping)
  - GridSphere for portal
  - Shibboleth (perhaps later) for federated trust models
  - Supporting technologies:
    - Tomcat
    - SQL Server
    - LDAP

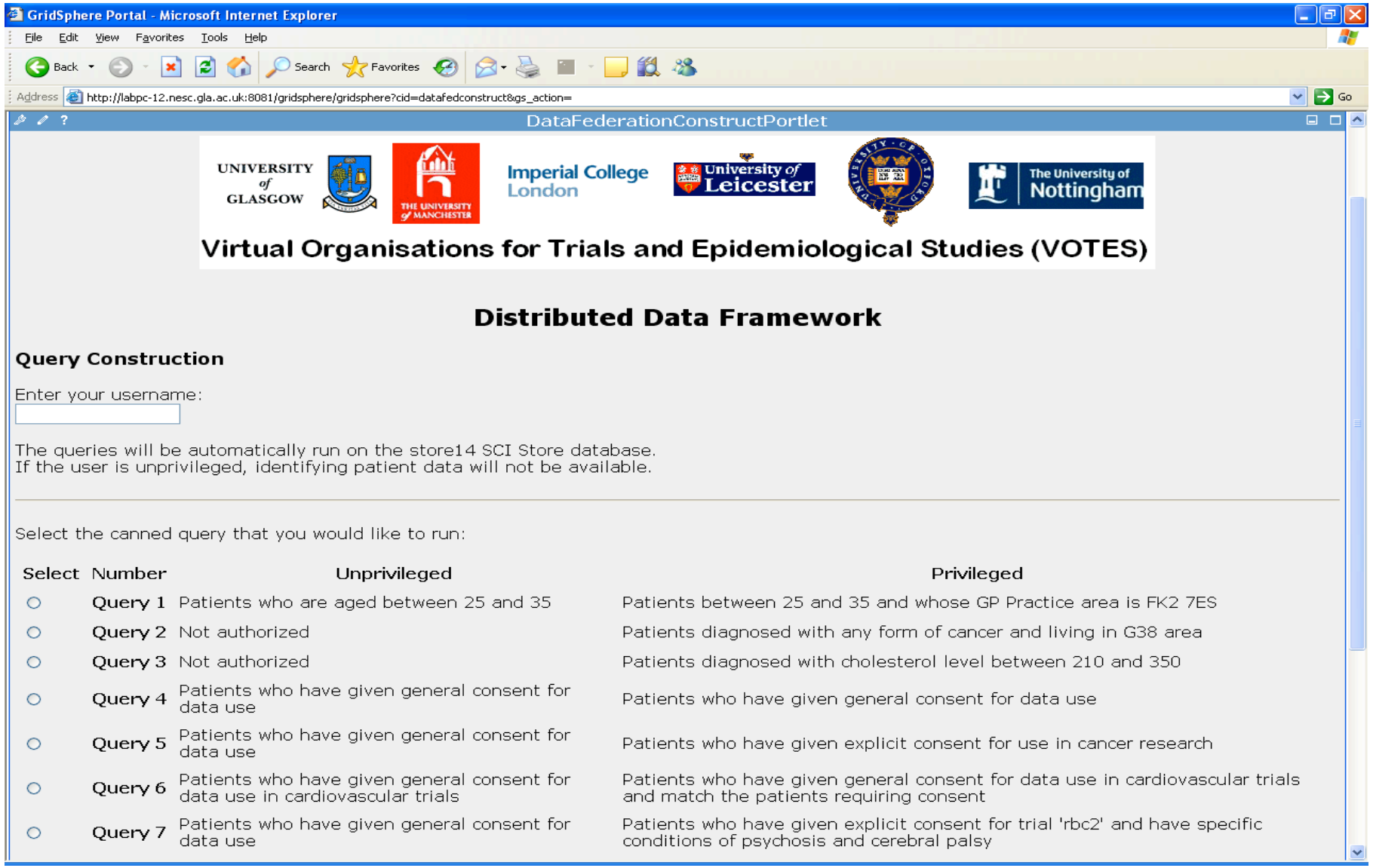
Copyright © 2006 University of Glasgow

Distributed Data Framework



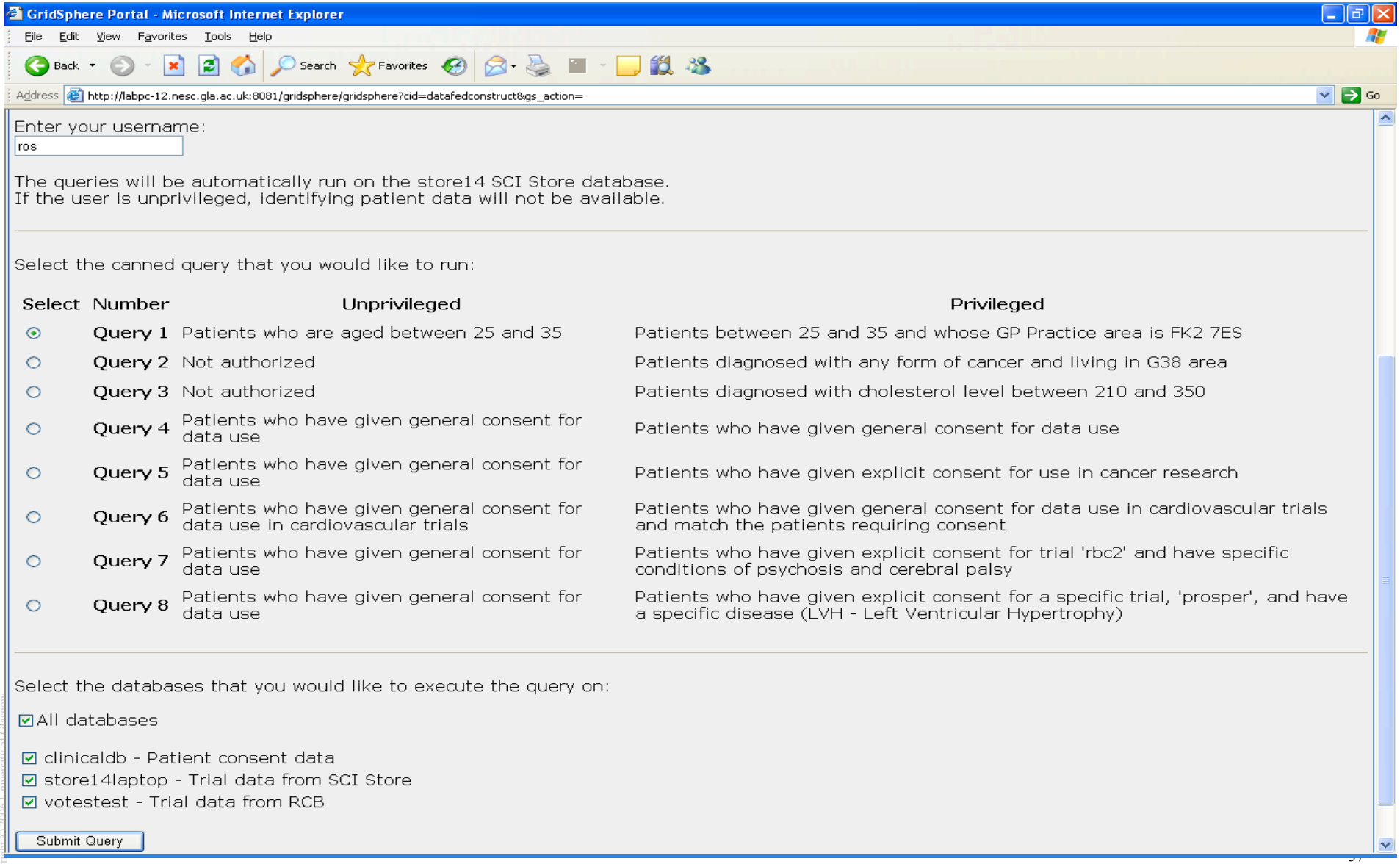
Copyright © 2006 University of Glasgow

Data Federation Portal
Alpha Prototype

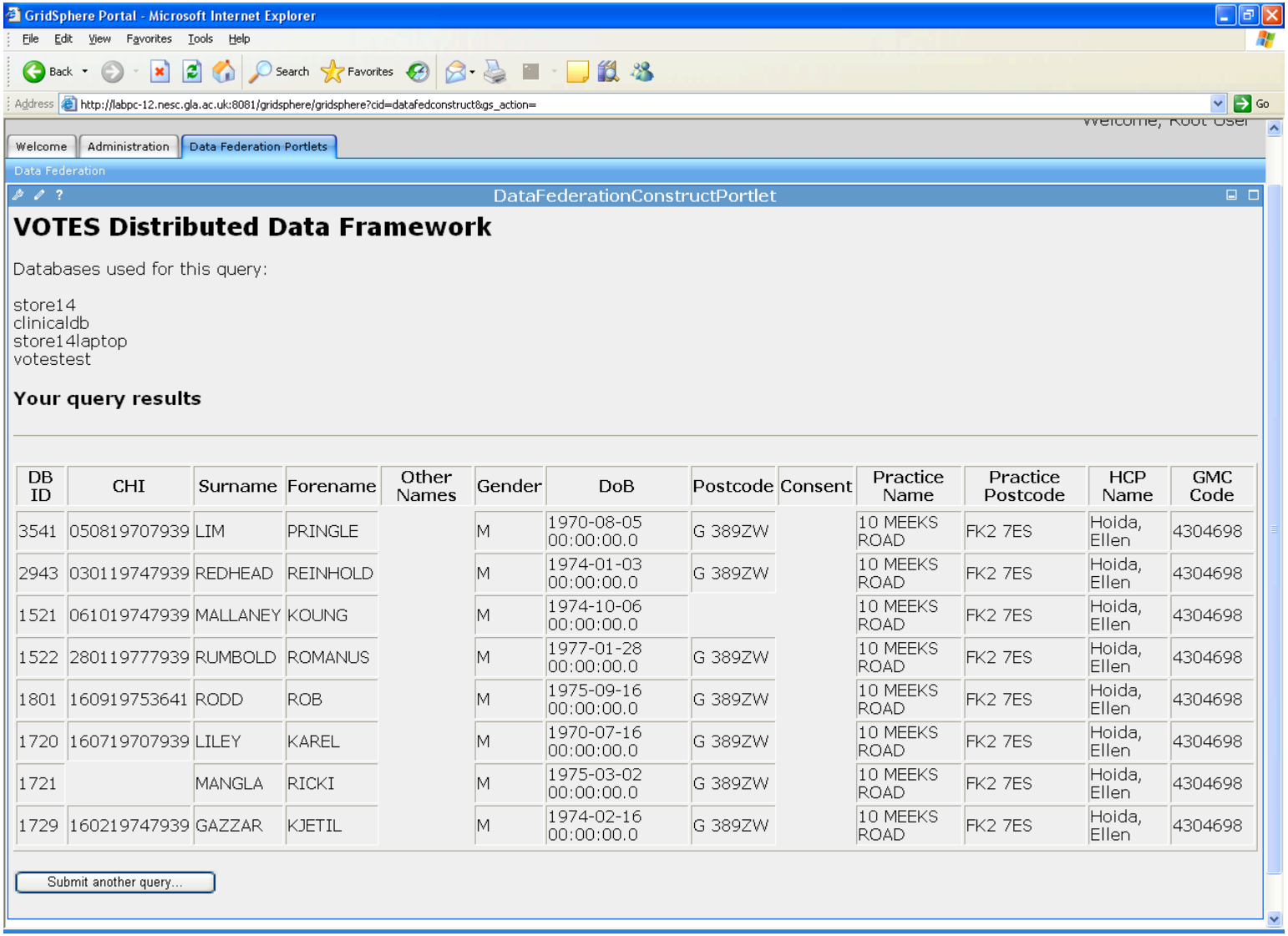


Copyright © 2006 University of Glasgow

Data Federation Portal
Alpha Prototype



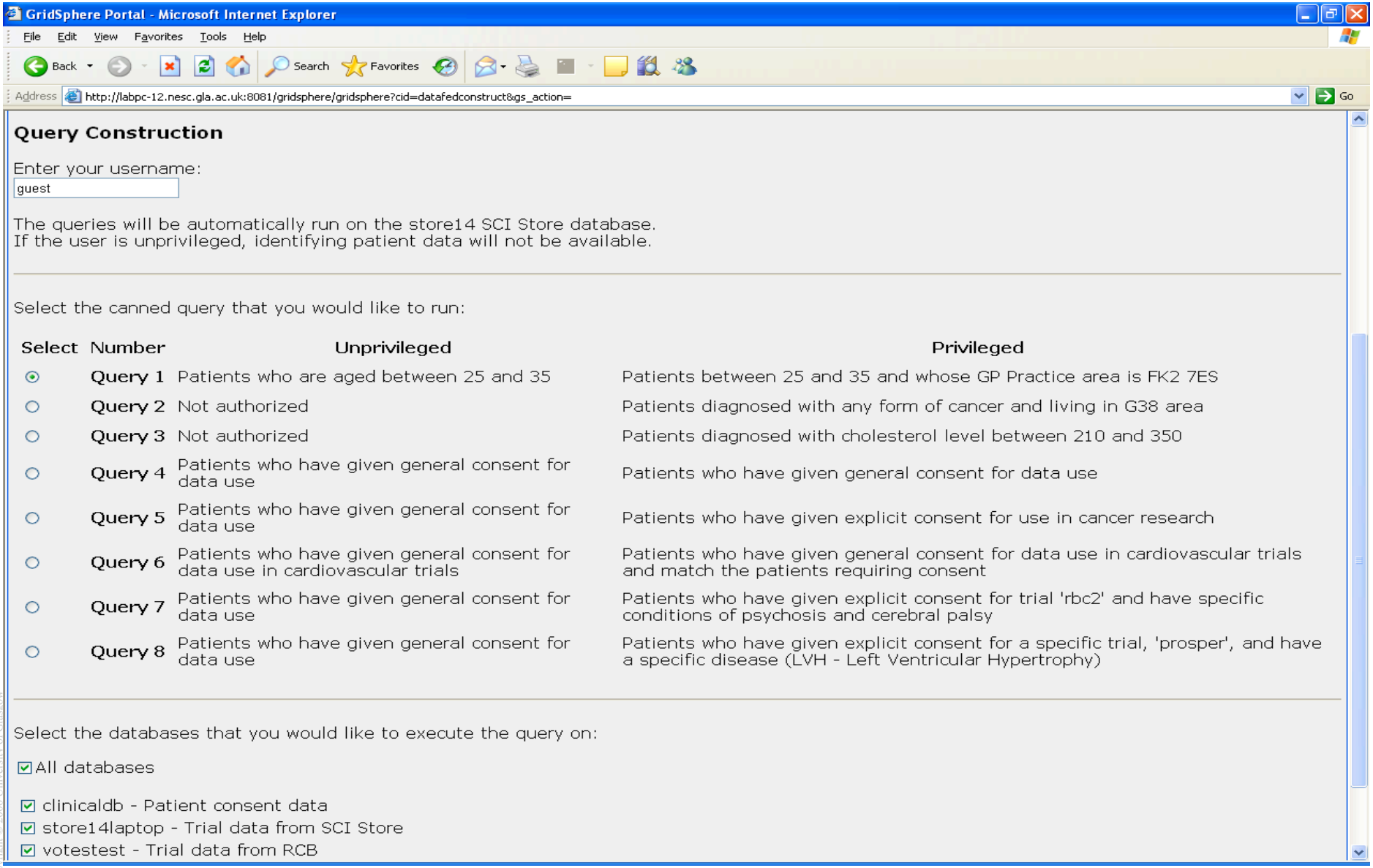
Data Federation Portal
Alpha Prototype



DB ID	CHI	Surname	Forename	Other Names	Gender	DoB	Postcode	Consent	Practice Name	Practice Postcode	HCP Name	GMC Code
3541	050819707939	LIM	PRINGLE		M	1970-08-05 00:00:00.0	G 389ZW		10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698
2943	030119747939	REDHEAD	REINHOLD		M	1974-01-03 00:00:00.0	G 389ZW		10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698
1521	061019747939	MALLANEY	KOUNG		M	1974-10-06 00:00:00.0			10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698
1522	280119777939	RUMBOLD	ROMANUS		M	1977-01-28 00:00:00.0	G 389ZW		10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698
1801	160919753641	RODD	ROB		M	1975-09-16 00:00:00.0	G 389ZW		10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698
1720	160719707939	LILEY	KAREL		M	1970-07-16 00:00:00.0	G 389ZW		10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698
1721		MANGLA	RICKI		M	1975-03-02 00:00:00.0	G 389ZW		10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698
1729	160219747939	GAZZAR	KJETIL		M	1974-02-16 00:00:00.0	G 389ZW		10 MEEKS ROAD	FK2 7ES	Hoida, Ellen	4304698

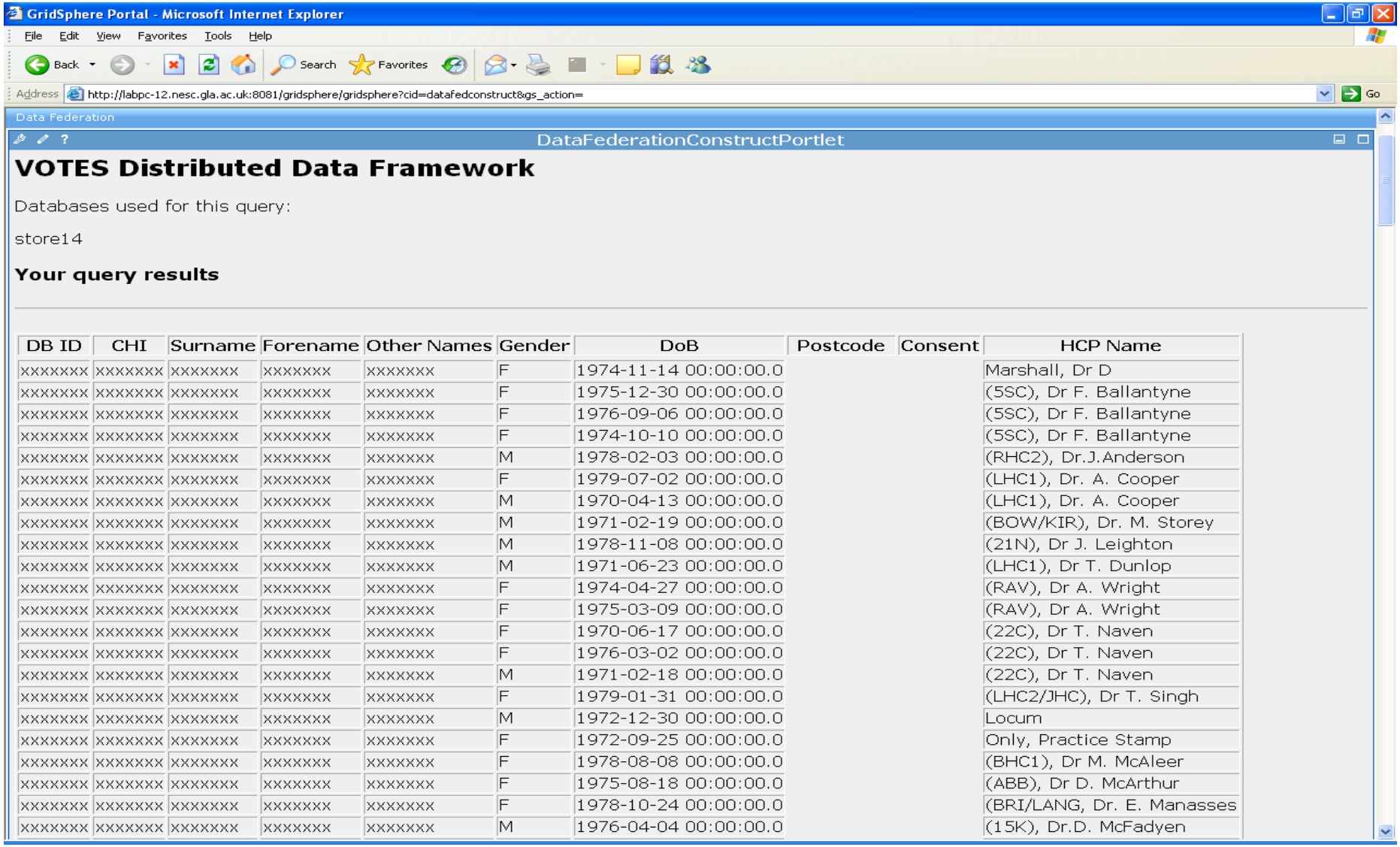
Copyright © 2006 University of Glasgow

Data Federation Portal
Alpha Prototype



Copyright © 2006 University of Glasgow

Data Federation Portal
Alpha Prototype



VOTES Distributed Data Framework

Databases used for this query:

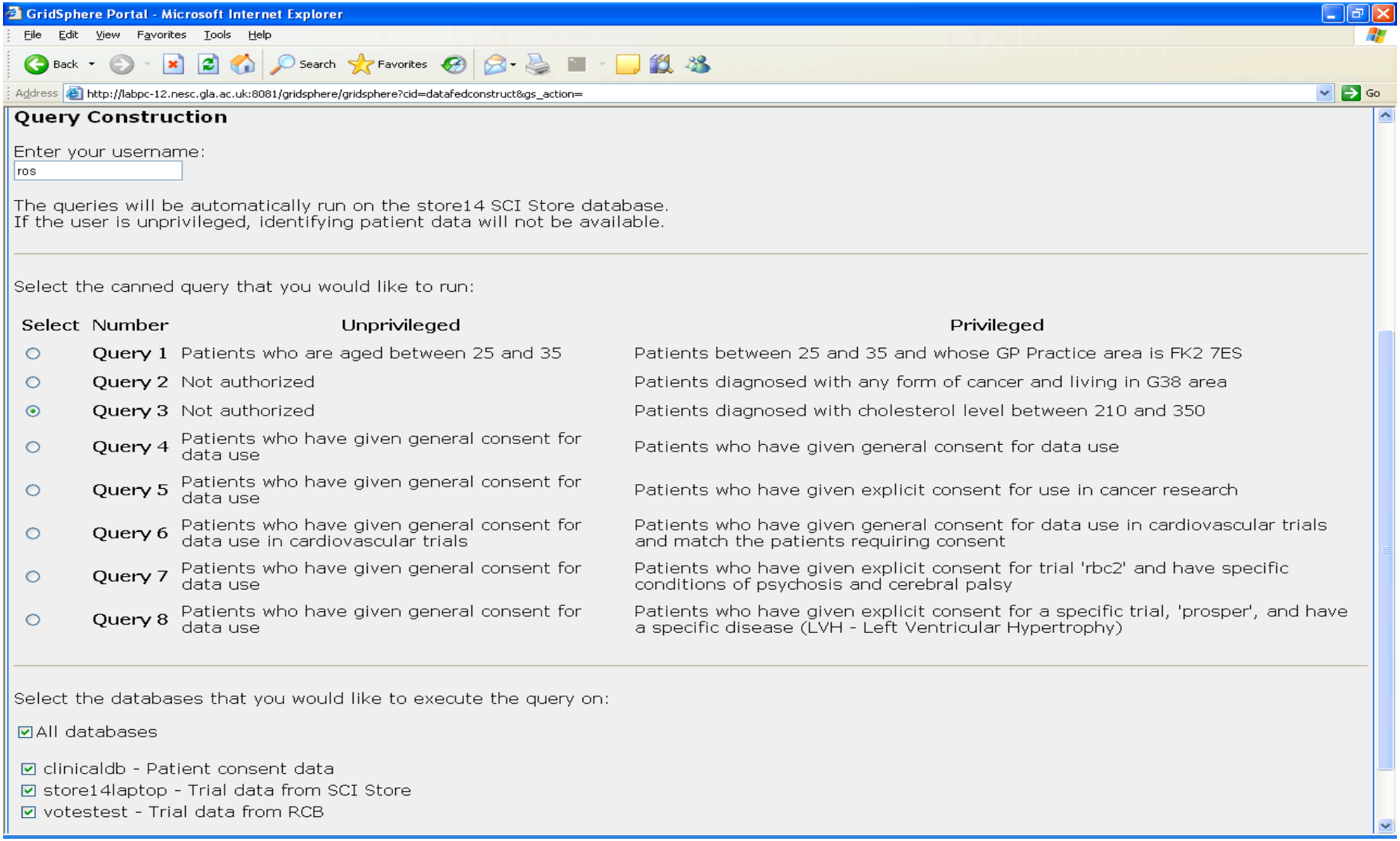
store14

Your query results

DB ID	CHI	Surname	Forename	Other Names	Gender	DoB	Postcode	Consent	HCP Name
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1974-11-14 00:00:00.0			Marshall, Dr D
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1975-12-30 00:00:00.0			(5SC), Dr F. Ballantyne
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1976-09-06 00:00:00.0			(5SC), Dr F. Ballantyne
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1974-10-10 00:00:00.0			(5SC), Dr F. Ballantyne
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1978-02-03 00:00:00.0			(RHC2), Dr.J.Anderson
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1979-07-02 00:00:00.0			(LHC1), Dr. A. Cooper
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1970-04-13 00:00:00.0			(LHC1), Dr. A. Cooper
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1971-02-19 00:00:00.0			(BOW/KIR), Dr. M. Storey
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1978-11-08 00:00:00.0			(21N), Dr J. Leighton
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1971-06-23 00:00:00.0			(LHC1), Dr T. Dunlop
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1974-04-27 00:00:00.0			(RAV), Dr A. Wright
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1975-03-09 00:00:00.0			(RAV), Dr A. Wright
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1970-06-17 00:00:00.0			(22C), Dr T. Naven
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1976-03-02 00:00:00.0			(22C), Dr T. Naven
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1971-02-18 00:00:00.0			(22C), Dr T. Naven
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1979-01-31 00:00:00.0			(LHC2/JHC), Dr T. Singh
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1972-12-30 00:00:00.0			Locum
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1972-09-25 00:00:00.0			Only, Practice Stamp
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1978-08-08 00:00:00.0			(BHC1), Dr M. McAleer
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1975-08-18 00:00:00.0			(ABB), Dr D. McArthur
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	F	1978-10-24 00:00:00.0			(BRI/LANG, Dr. E. Manasses
xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	xxxxxxx	M	1976-04-04 00:00:00.0			(15K), Dr.D. McFadyen

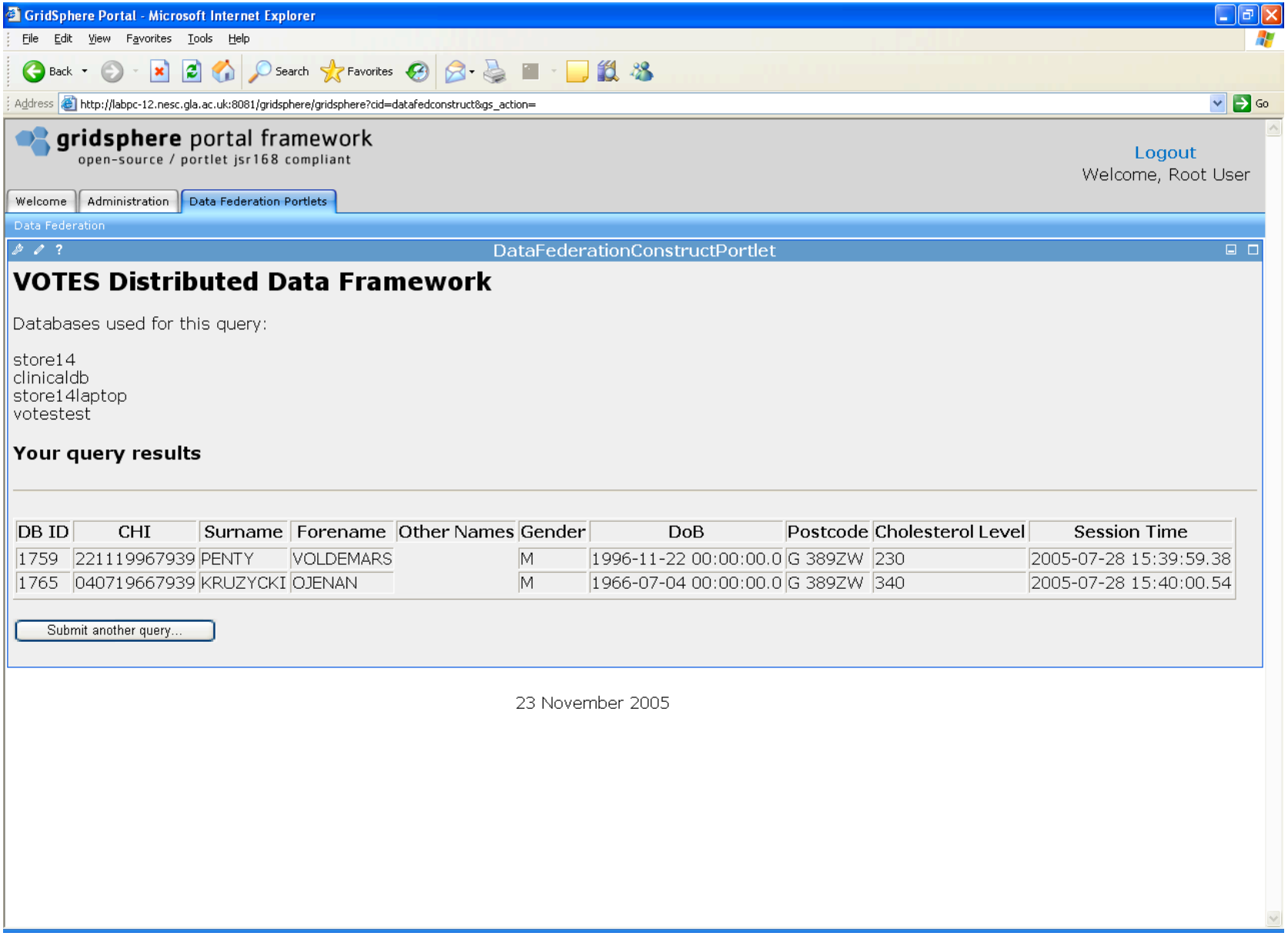
Copyright © 2006 University of Glasgow

Data Federation Portal
Alpha Prototype



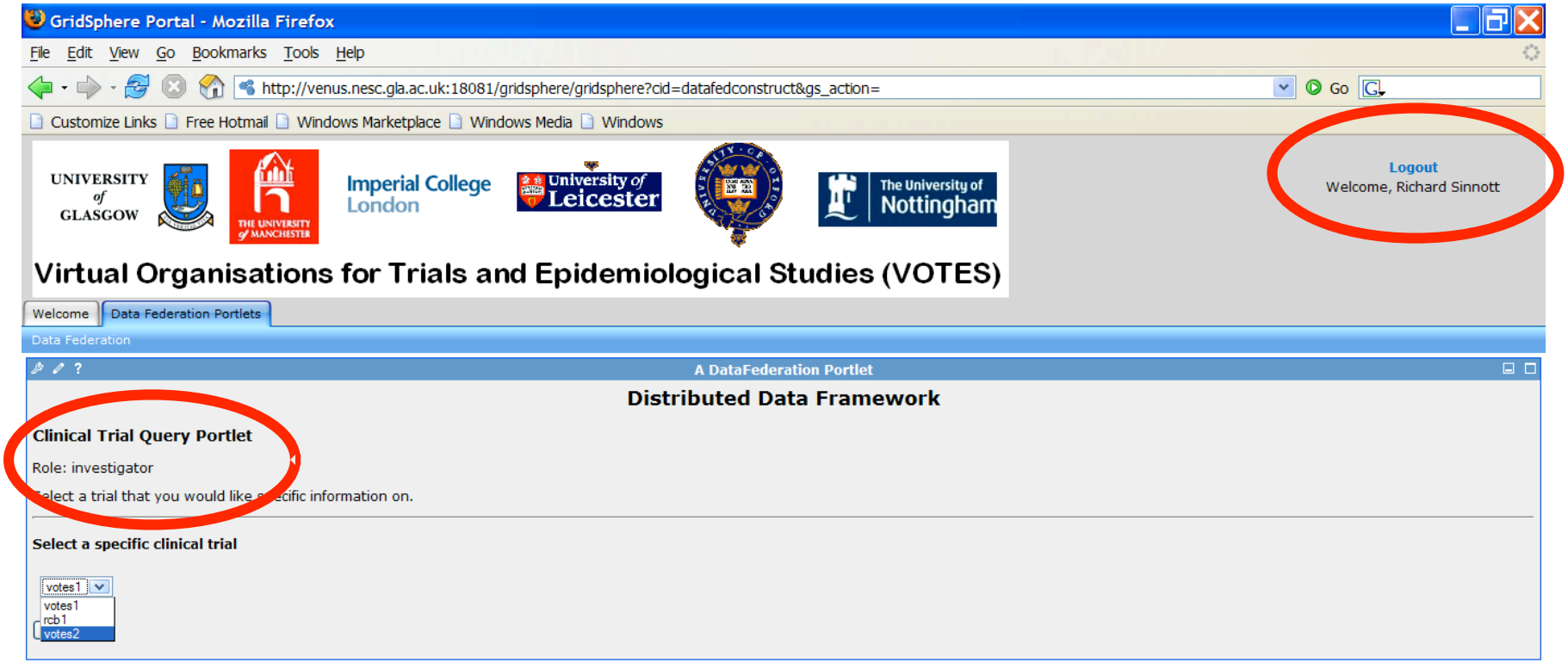
Copyright © 2006 University of Glasgow

Data Federation Portal
Alpha Prototype



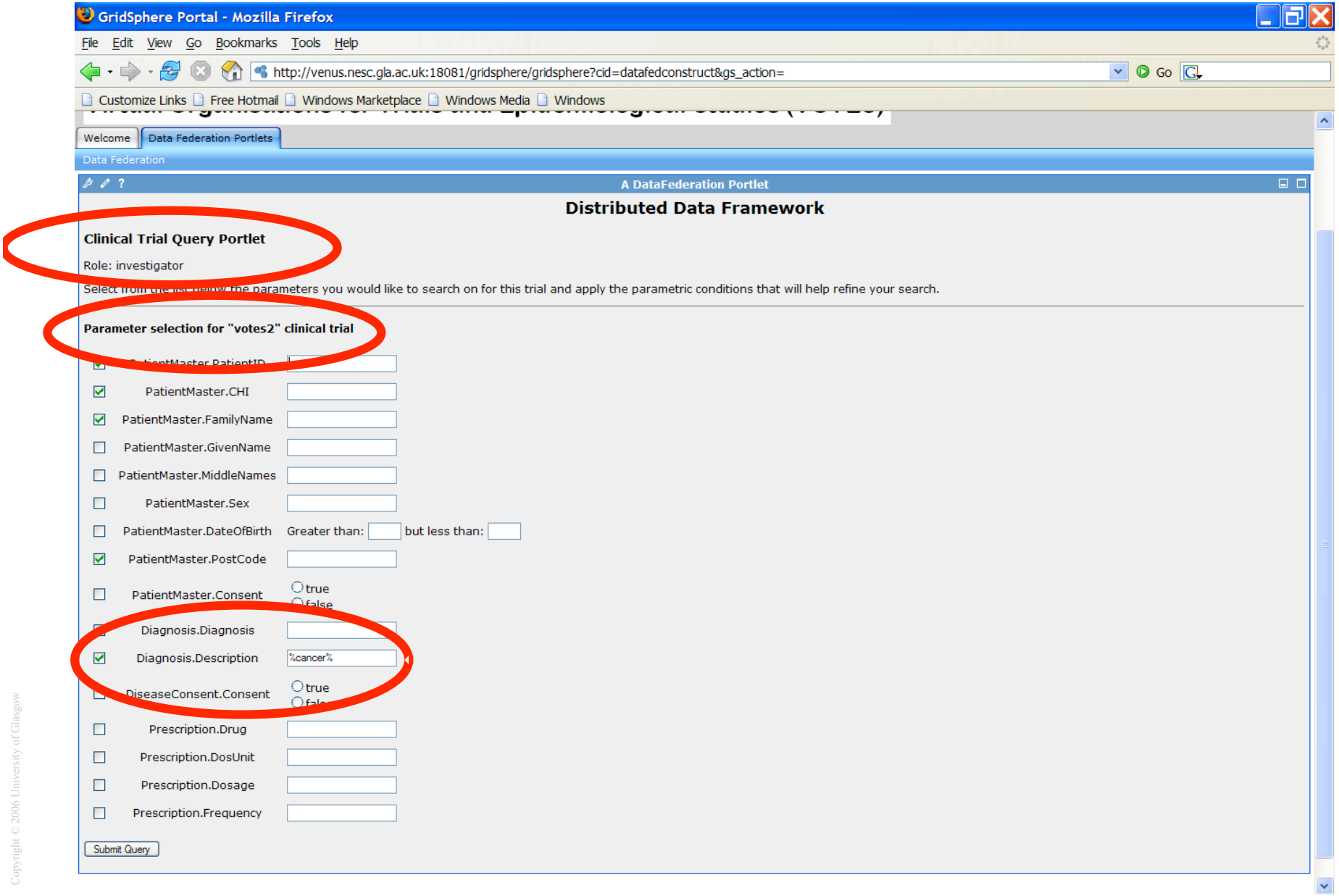
Copyright © 2006 University of Glasgow

Data Federation Portal
Beta Prototype



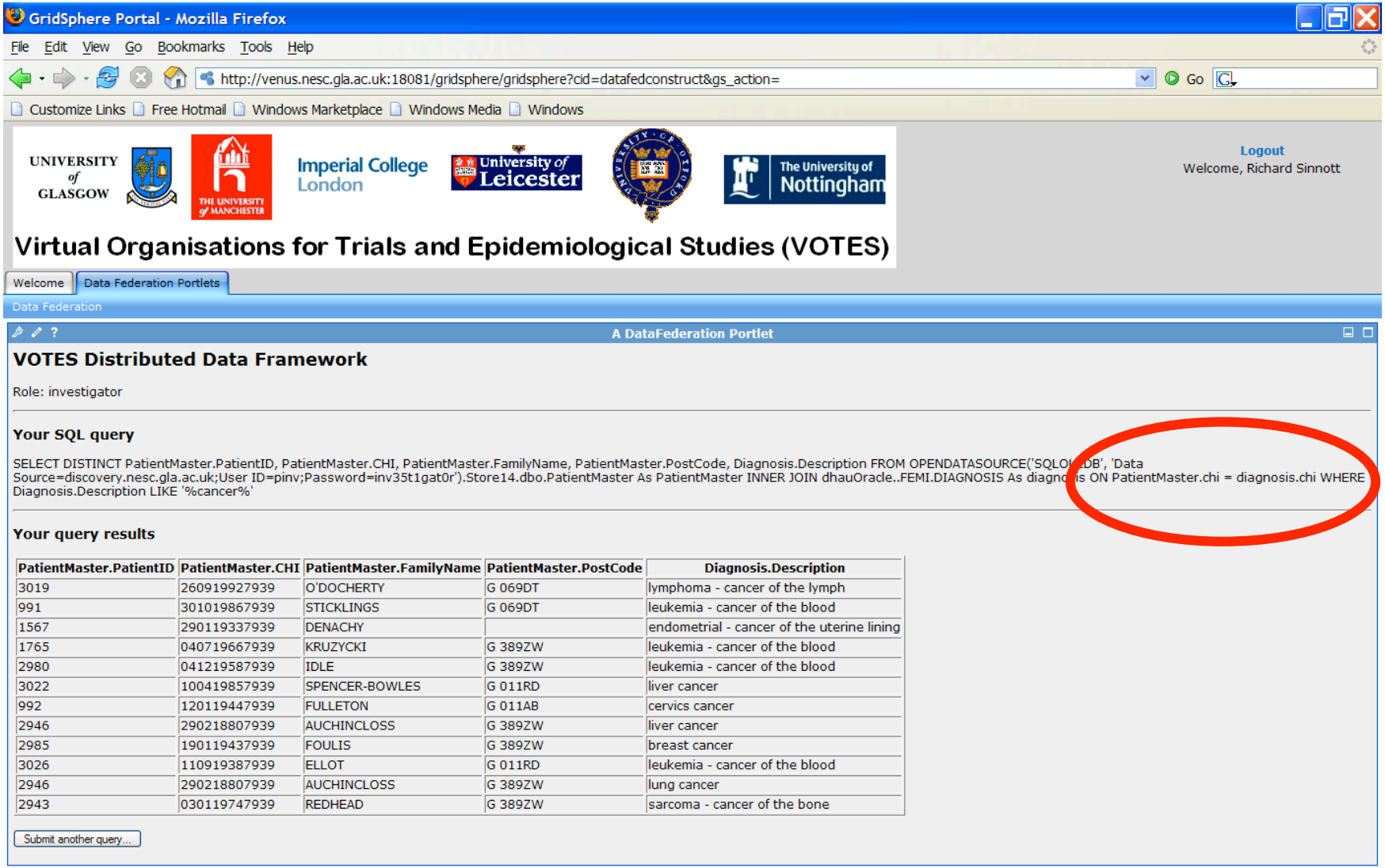
07 March 2006

Data Federation Portal
Beta Prototype



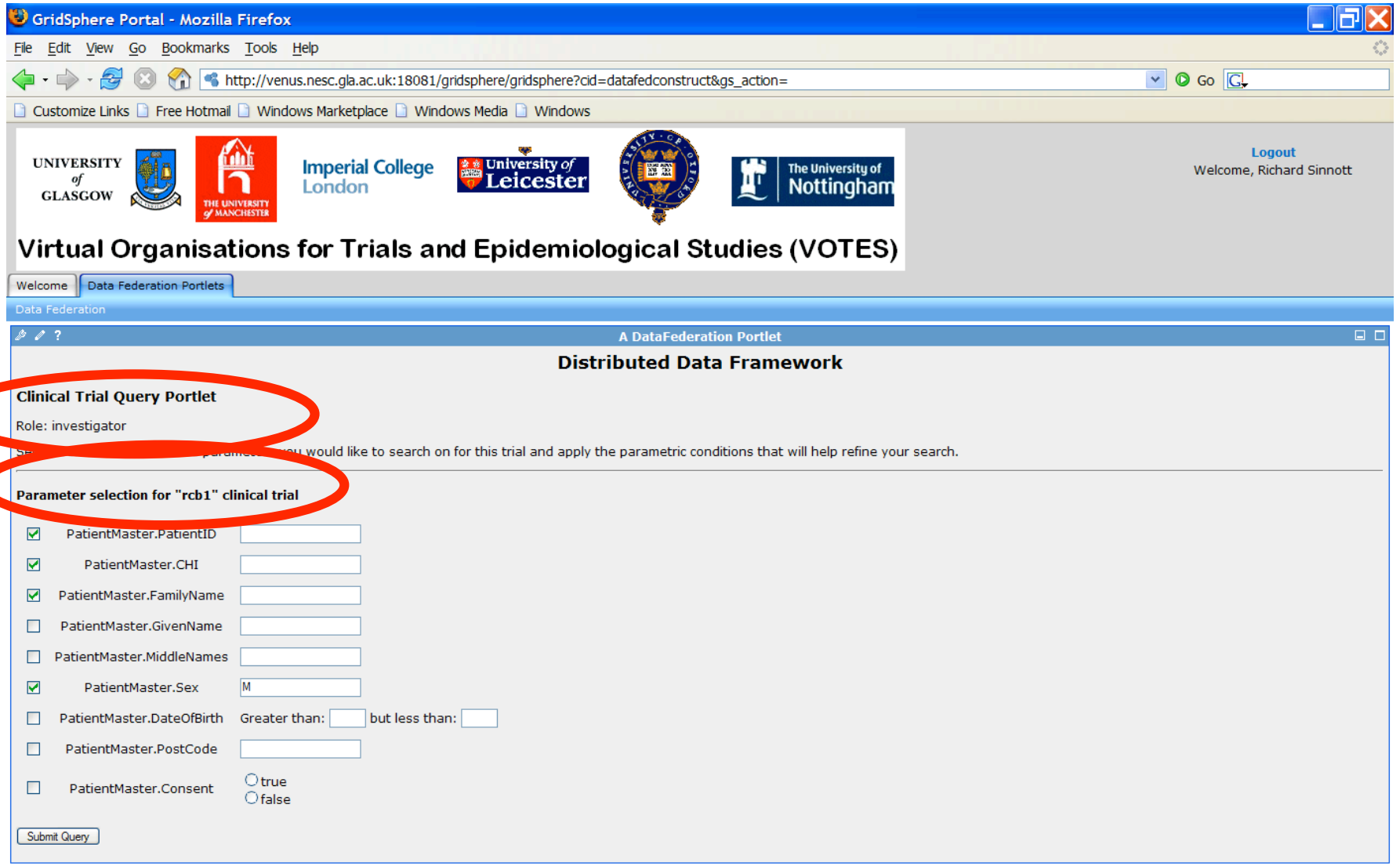
Copyright © 2006 University of Glasgow

Data Federation Portal
Beta Prototype



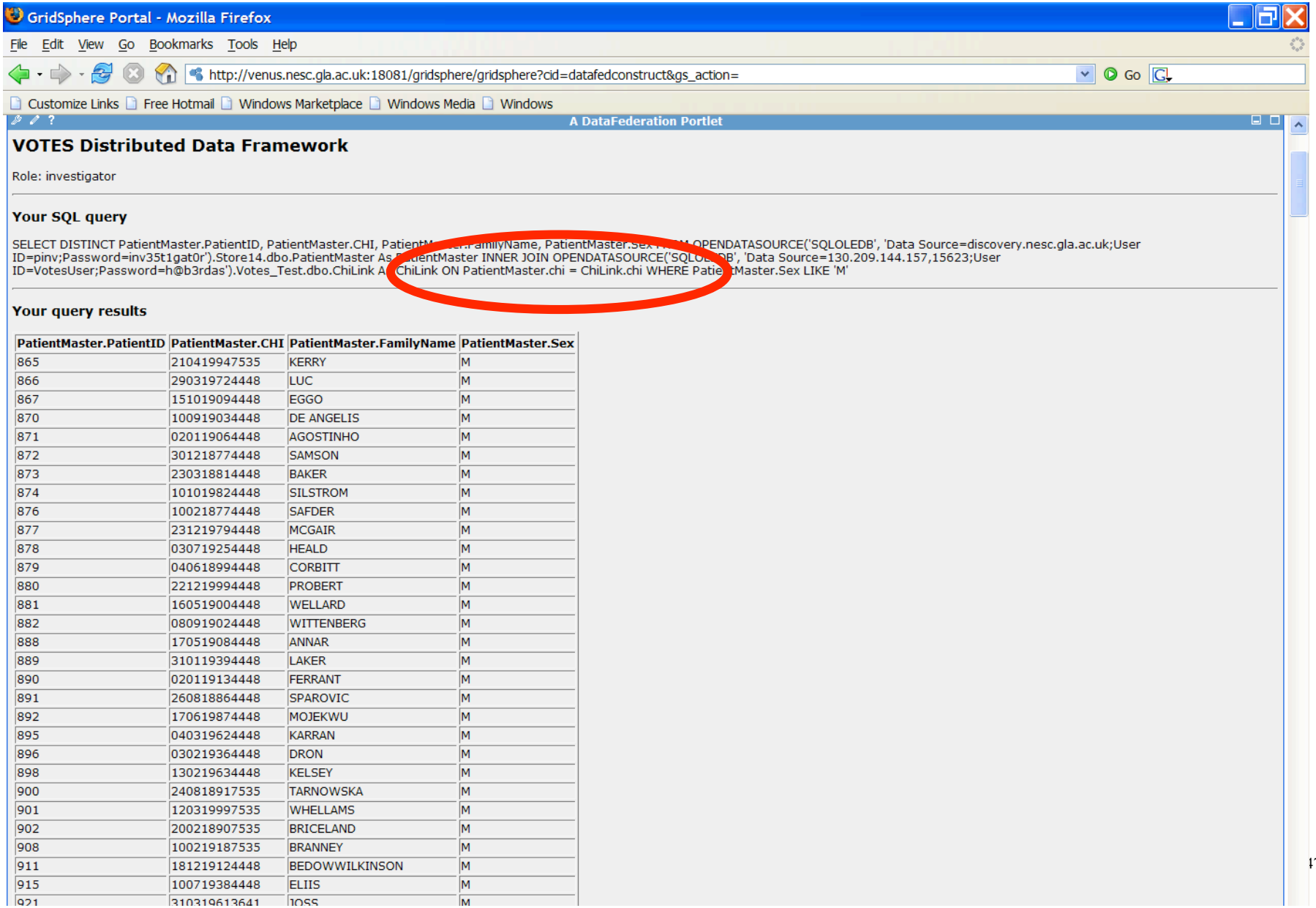
Copyright © 2006 University of Glasgow

Data Federation Portal
Beta Prototype



Copyright © 2006 University of Glasgow

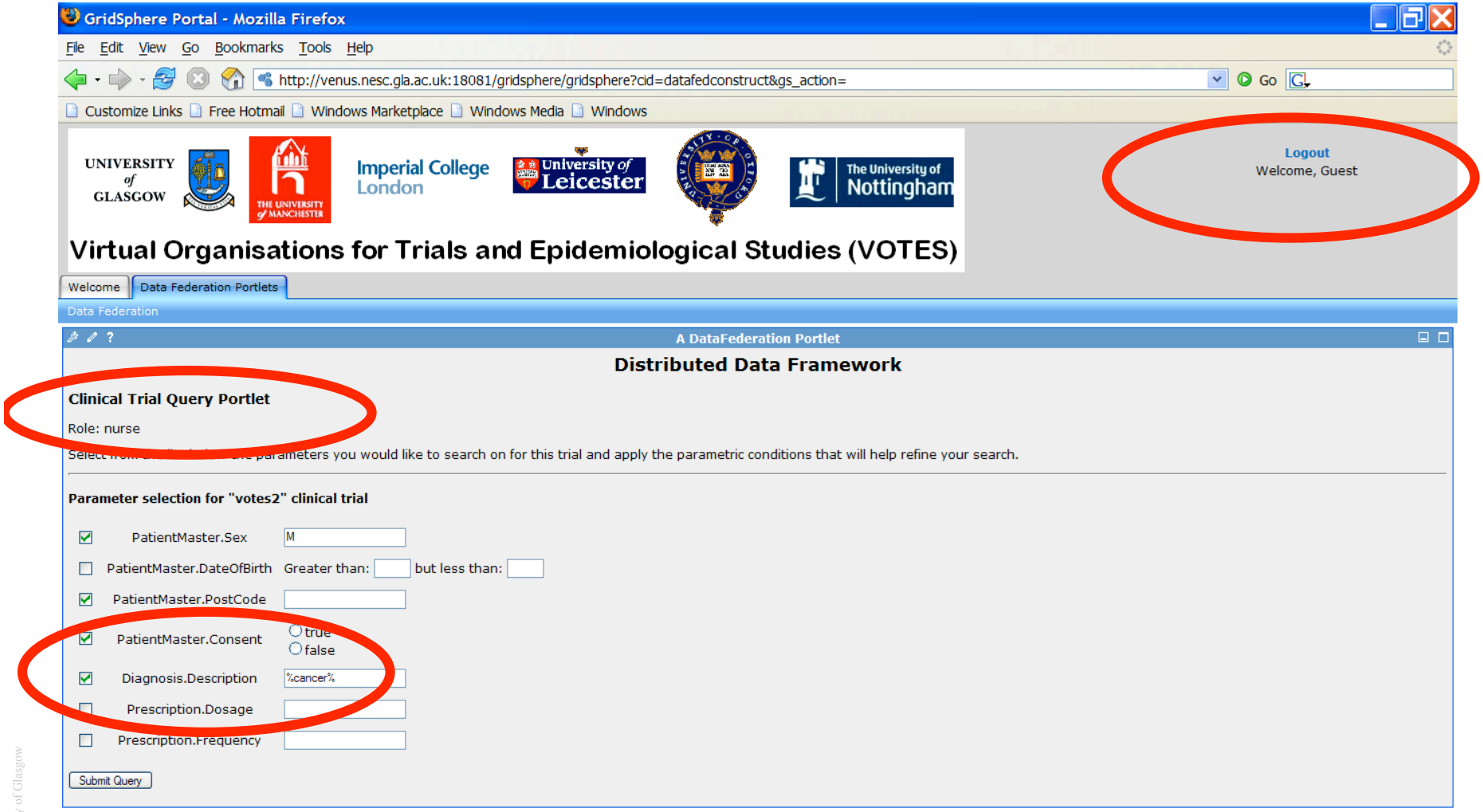
Data Federation Portal
Beta Prototype



PatientMaster.PatientID	PatientMaster.CHI	PatientMaster.FamilyName	PatientMaster.Sex
865	210419947535	KERRY	M
866	290319724448	LUC	M
867	151019094448	EGGO	M
870	100919034448	DE ANGELIS	M
871	020119064448	AGOSTINHO	M
872	301218774448	SAMSON	M
873	230318814448	BAKER	M
874	101019824448	SILSTROM	M
876	100218774448	SAFDER	M
877	231219794448	MCGAIR	M
878	030719254448	HEALD	M
879	040618994448	CORBITT	M
880	221219994448	PROBERT	M
881	160519004448	WELLARD	M
882	080919024448	WITTENBERG	M
888	170519084448	ANNAR	M
889	310119394448	LAKER	M
890	020119134448	FERRANT	M
891	260818864448	SPAROVIC	M
892	170619874448	MOJEKWU	M
895	040319624448	KARRAN	M
896	030219364448	DRON	M
898	130219634448	KELSEY	M
900	240818917535	TARNOWSKA	M
901	120319997535	WHELLAMS	M
902	200218907535	BRICELAND	M
908	100219187535	BRANNEY	M
911	181219124448	BEDOWWILKINSON	M
915	100719384448	ELIIS	M
921	310319613641	JOSS	M

Copyright © 2006 University of Glasgow

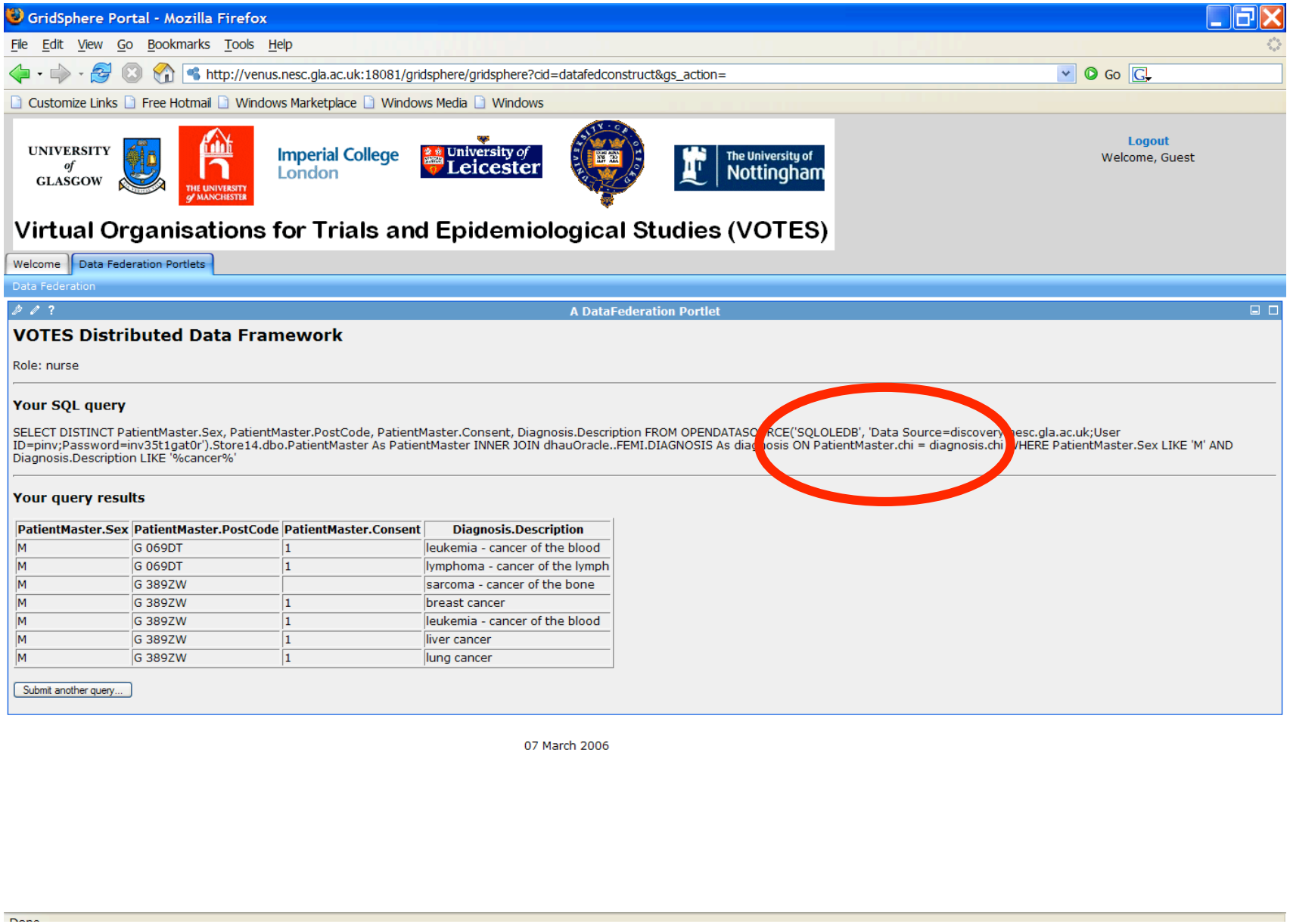
Data Federation Portal
Beta Prototype



07 March 2006

Copyright © 2006 University of Glasgow

Data Federation Portal
Beta Prototype



07 March 2006

Copyright © 2006 University of Glasgow

Conclusions

- Numerous application domains exploring e-Science/Grid technologies
- Consolidation of know-how/technologies essential
  - EGEE
  - OMII
  - UK e-Science task forces (ETF, STF, ATF, …)
  - NDCC
  - NeSC
- Do we know how best to build Grids?
  - Different domains coming up with own ways of building Grids
    - OGSA needed asap
  - Clear that various domains have issues which must be resolved before Grid can make significant **and long lasting** impact
- NeSC Glasgow is leader in security/life sciences but still lots to do/learn…

Copyright © 2006 University of Glasgow