INWA : using OGSA-DAI between the UK, Australia and China

Terry Sloan

EPCC, The University of Edinburgh

t.sloan@epcc.ed.ac.uk

Overview

- The Grid vision
- The INWA project
- Experiences from data mining over the grid OGSA-DAI
- Typical scenario
- Barriers
- Future Plans

The Grid Vision

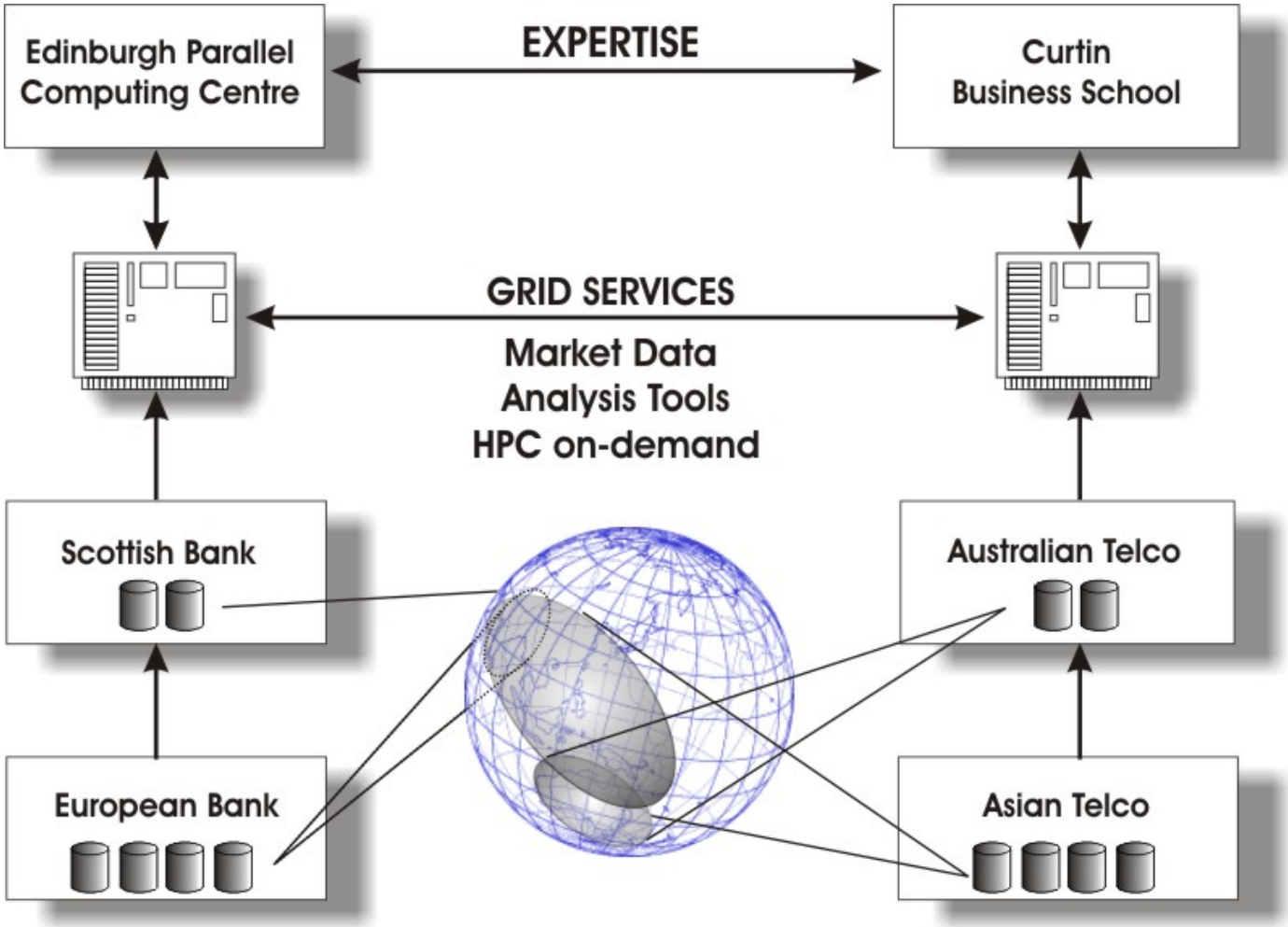
"… flexible, secure, coordinated resource sharing among dynamic collections of individuals, institutions and resources - what we refer to as virtual organisations."

The Anatomy of the Grid: Enabling Scalable Virtual Organizations. I. Foster, C. Kesselman, S. Tuecke. *International J. Supercomputer Applications*, 15(3), 2001.

epcc

The INWA Project

The INWA virtual organisation



INWA Resources & Participants

- Resources
  - UK mortgage data
  - UK property data
  - Australian telco data
  - Australian property data
  - Compute power at EPCC
  - Compute power at Curtin

- Individuals and Organisations:
  - Analyst at EPCC, UK
  - Analyst at Curtin, Australia
  - EPCC, UK – compute resource provider and host
  - Curtin, Australia – compute resource host
  - Sun Microsystems, Aus – compute resource provider
  - Bank, UK – data provider
  - ESPC, UK – data provider
  - Telco, Aus – data provider
  - VGO, WA, Aus – data provider

Background

- Funded by UK Economic & Social Research Council (UK) in the Pilot Projects in E-Social Science
  - Small scale projects to explore the potential of Grid technologies within the social sciences
  - Informing Business & Regional Policy: Grid enabled fusion of global data & local knowledge
  - INWA : Innovation Node Western Australia
- Started November 2003
  - Initial phase finished August 2004

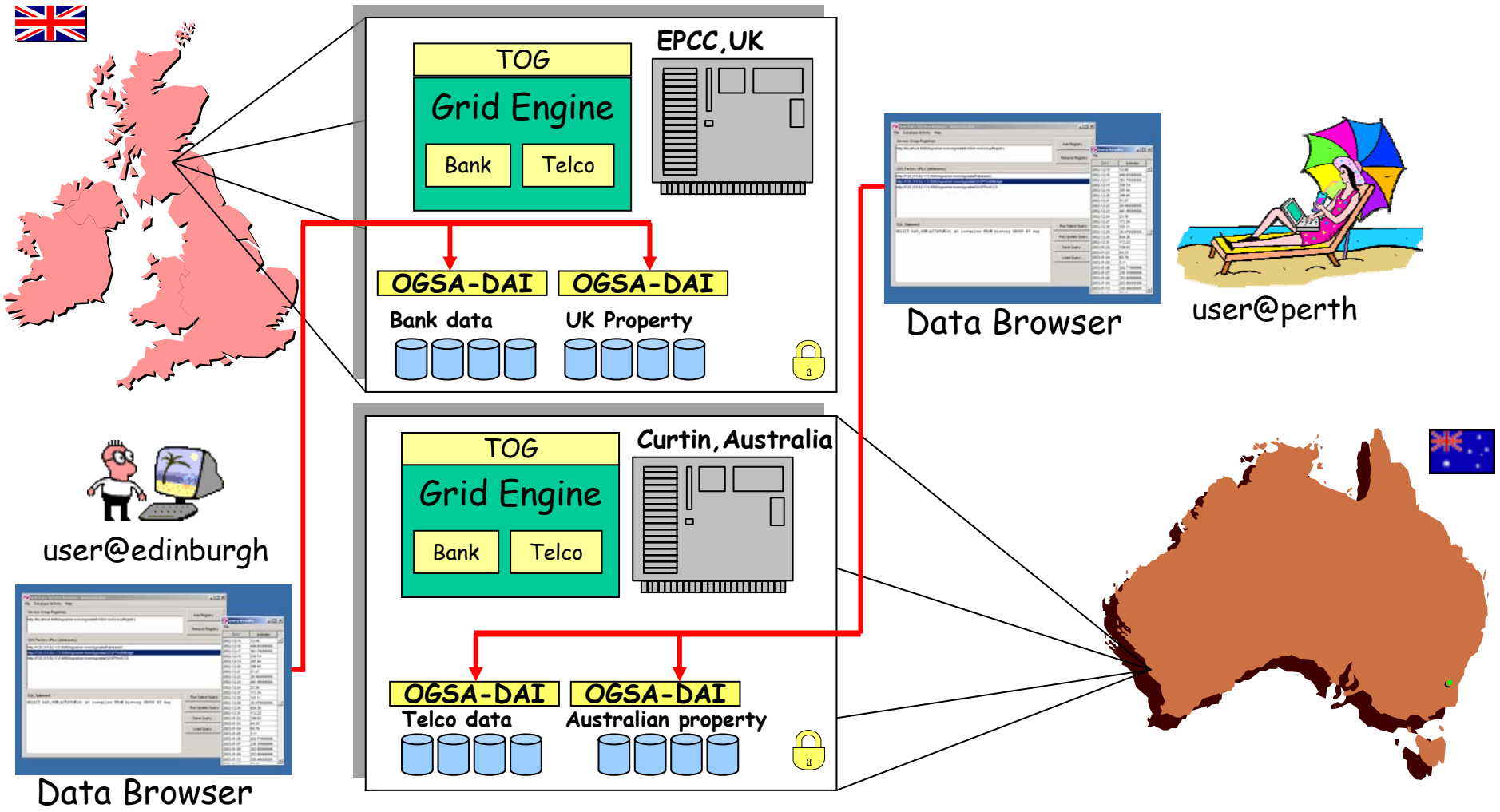
Project Aims

- Evaluate the suitability of existing grid solutions for secure distributed data mining and analysis on commercially sensitive data
- Investigate the advantages of fusing public and private data enabled by a grid environment

Barriers to Success

- Can existing grid technologies fulfill this vision?
  - Transfer-queue Over Globus (TOG) v1.1 from the UK e-Science Sun Data and Compute Grids project
    - provides access to remote HPC resource
  - Open Grid Services Architecture – Data Access and Integration (OGSA-DAI) Release 3.1
    - provides access control and discovery of distributed heterogeneous data resources
  - First Data Investigation on the Grid (FirstDIG)
    - grid data service browser provides SQL access to OGSA-DAI enabled resources
    - now part of OGSA-DAI R4.0
  - Globus Toolkit 2 and 3
    - Grid middleware
- If not what are the barriers?
  - Technology?
  - Socio-economic?

The INWA Grid



epcc

Data Mining over the Grid

Data mining

- A typical data mining project broadly involves
  1. Getting the data
  2. Cleaning it
  3. Mining it
- Iteration through steps 1 to 3 to refine models
- So where can the Grid help?

Getting the data

- Traditionally a file export
  - But OGSA-DAI is available
    - Open Grid Services Architecture : Data Access and Integration
    - Assists with the access and integration of data from separate data sources via the Grid
  - But organisations will not contemplate external access to operational/sensitive data
  - So back to a file export
- UK Land registry
  - Public data source but no OGSA-DAI interface
  - Appropriate mechanisms need to be in place before data sharing can take place
- So simulated this access over the Grid
  - But some security issues

Data Fusion

- Fusing commercial data with public property data

Account ID	*Address*	*Loan*	*Date*	…
2289738	10 Downing Street, …	200,000	10/2/2002	…
2672623	20 My Street, …	100,000	14/8/1980	…

+

Address	*#Bedrooms*	*#Garages*	…
10 Downing Street, …	4	3	…
20 My Street, …	3	0	…

=

Account ID	*Address*	*Loan*	*Date*	*#Bedrooms*	*#Garages*	…
2289738	10 Downing …	200,000	10/2/2002	4	3	…
2672623	20 My Street, …	100,000	14/8/1980	3	0	…

Data Fusion

- Why do it ?
  - Prospect of better models/predictions
  - Added value
- But
  - need a distributed-aggregated approach to preserve anonymity
- So simulated this over the Grid
  - Using a less specific join key
    - Not a 1-1 join but a 1-n so averaging necessary
  - Limited the potential gains from fusion
- Fuzzy joins
  - e.g. postcode formats, addresses (St=Street, flat numbers)

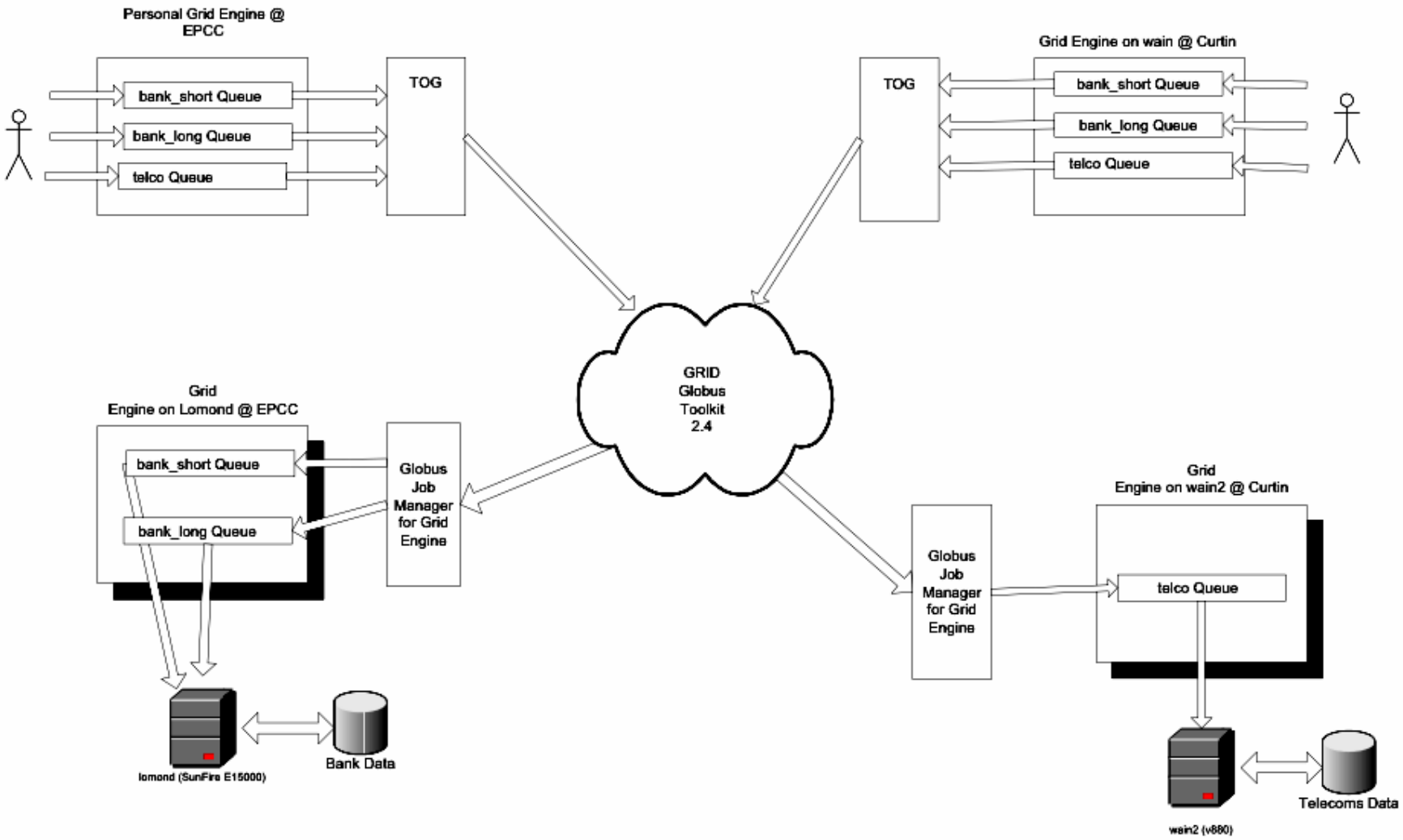
Data Fusion tool support

- Little real support for data integration over the Grid
  - OGSA-DQP (Distributed Query Processing) is limited
    - Needs Linux and so is restrictive
    - Uses OQL which similar to SQL but not as common
    - Complicated set-up
    - Dependent on a number of nodes being available to provide services
- Used FirstDIG browser
  - Relevant data pulled over
  - Data joined locally
  - This works but obviously is not ideal
    - A lot of user interaction is required.
      - 7 queries are necessary to join two datasets
- So again limited success over the Grid

Grid Computation

- Large data sets so, …
- Cleaning and mining jobs sent to where data is resident (UK and Australia)
- Globus Toolkit V2.x (GT2), Grid Engine and TOG used
- But…
  - Installation issues with GT2
    - Not out-of-the-box, requires significant time, effort, expertise
  - Security issues with GT2 & TOG
    - Bug in the Globus Java CoG Kit
    - Security flag omission in TOG
- All now works and is currently being used between UK and Australia

TOG/GridEngine/Globus set-up



Typical scenario

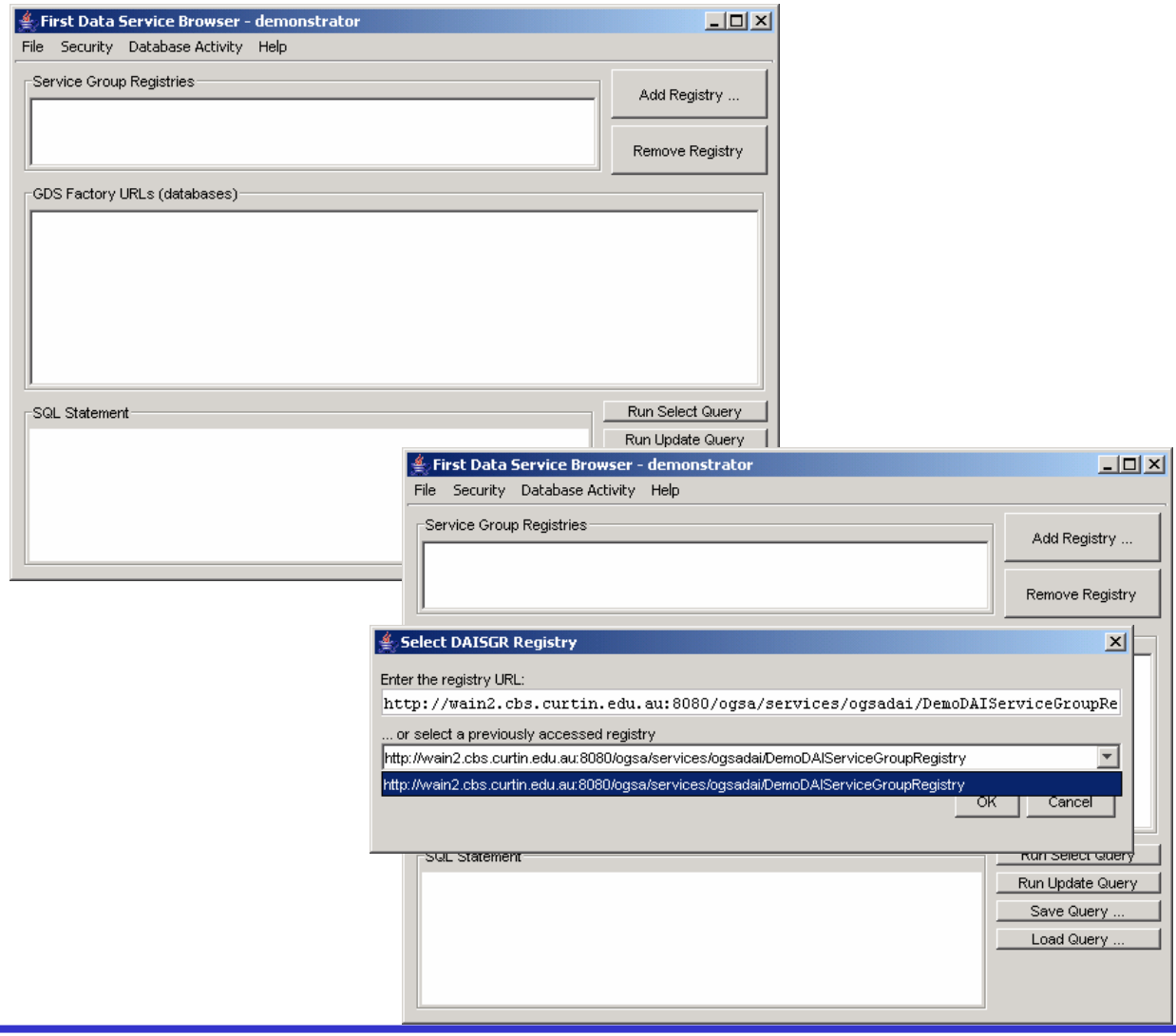
Demonstration

Scenario

- A bank wants to predict if home owners are likely to move house within 5 years of taking out a loan to buy the house
- This type of loan is a mortgage
- Bank wants to use its own data and publically available data to help improve the prediction
- Demo uses dummy data
- Data stored in Australia in OGSA-DAI enabled databases
- Demo shows an example of a workflow used in the project to browse and analyse data
- FirstDIG browser and OGSA-DAI were used to browse and fuse data

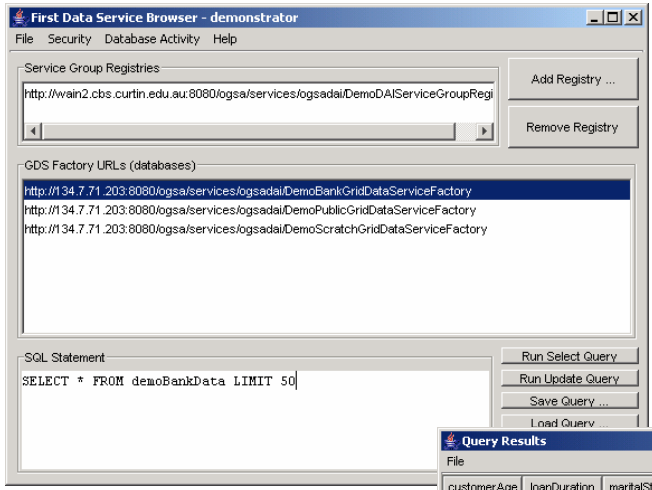
Access OGSA-DAI Registry

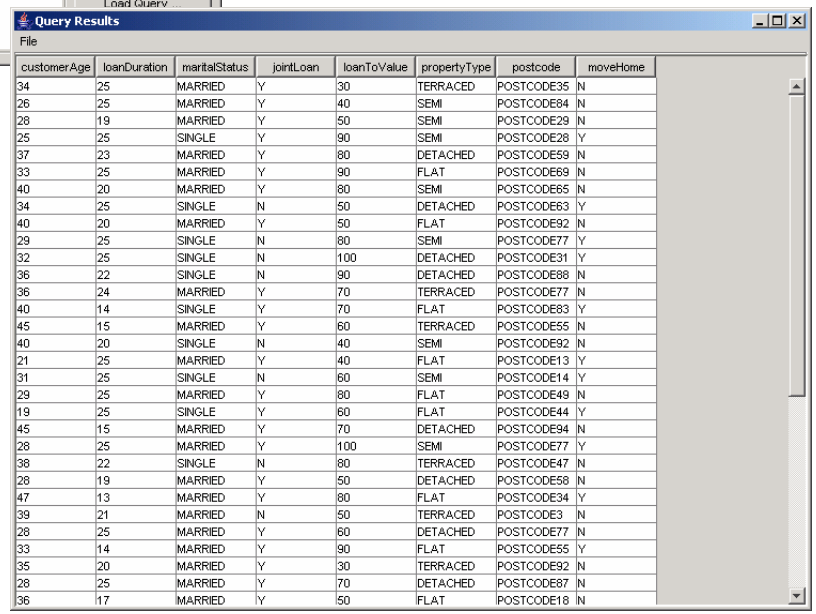
- FirstDIG browser started
- OGSA-DAI registry at Curtin selected
  - Data sources available



Browse demo bank data

- Grid data service factories appear
- demoBank GDSF selected
- SQL query input
  - select * from demoBankData LIMIT 50
- Run select query
- Query results appear
  - example bank data



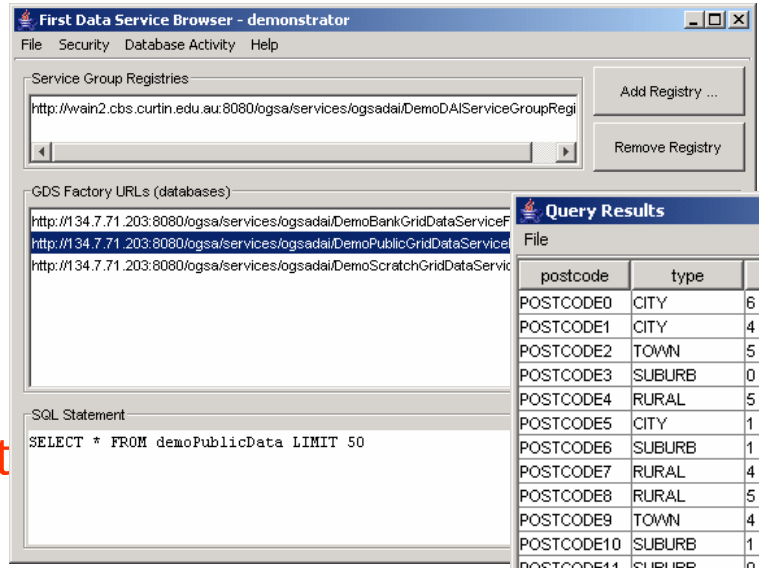

Query Results

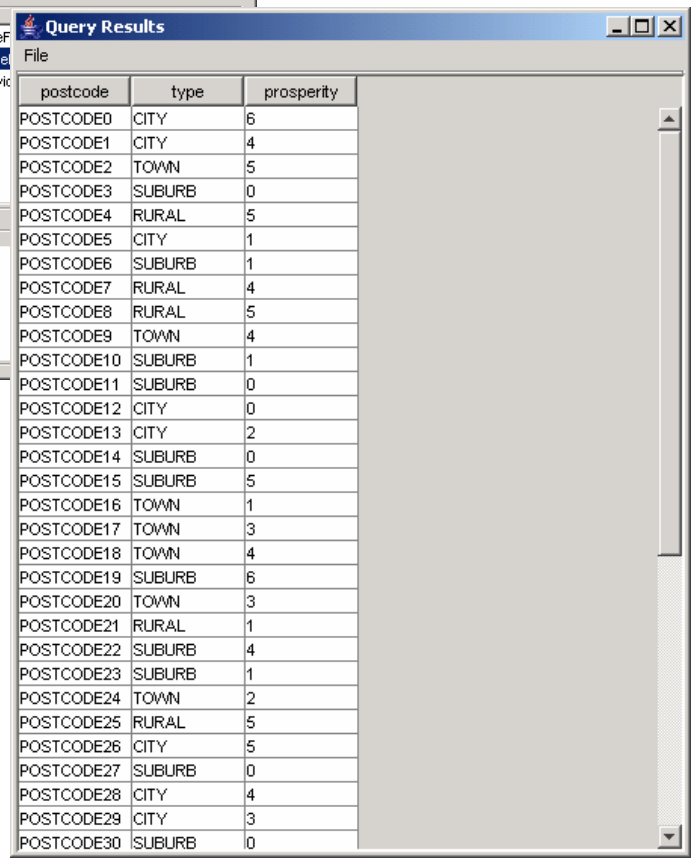
File

customerAge	loanDuration	maritalStatus	jointLoan	loanToValue	propertyType	postcode	moveHome
34	25	MARRIED	Y	30	TERRACED	POSTCODE35	N
26	25	MARRIED	Y	40	SEMI	POSTCODE84	N
28	19	MARRIED	Y	50	SEMI	POSTCODE29	N
25	25	SINGLE	Y	90	SEMI	POSTCODE28	Y
37	23	MARRIED	Y	80	DETACHED	POSTCODE59	N
33	25	MARRIED	Y	90	FLAT	POSTCODE69	N
40	20	MARRIED	Y	80	SEMI	POSTCODE65	N
34	25	SINGLE	N	50	DETACHED	POSTCODE63	Y
40	20	MARRIED	Y	50	FLAT	POSTCODE92	N
29	25	SINGLE	N	80	SEMI	POSTCODE77	Y
32	25	SINGLE	N	100	DETACHED	POSTCODE31	Y
36	22	SINGLE	N	90	DETACHED	POSTCODE88	N
36	24	MARRIED	Y	70	TERRACED	POSTCODE77	N
40	14	SINGLE	Y	70	FLAT	POSTCODE83	Y
45	15	MARRIED	Y	60	TERRACED	POSTCODE55	N
40	20	SINGLE	N	40	SEMI	POSTCODE92	N
21	25	MARRIED	Y	40	FLAT	POSTCODE13	Y
31	25	SINGLE	N	60	SEMI	POSTCODE14	Y
29	25	MARRIED	Y	80	FLAT	POSTCODE49	N
19	25	SINGLE	Y	60	FLAT	POSTCODE44	Y
45	15	MARRIED	Y	70	DETACHED	POSTCODE94	N
28	25	MARRIED	Y	100	SEMI	POSTCODE77	Y
38	22	SINGLE	N	80	TERRACED	POSTCODE47	N
28	19	MARRIED	Y	50	DETACHED	POSTCODE58	N
47	13	MARRIED	Y	80	FLAT	POSTCODE34	Y
39	21	MARRIED	N	50	TERRACED	POSTCODE3	N
28	25	MARRIED	Y	60	DETACHED	POSTCODE77	N
33	14	MARRIED	Y	90	FLAT	POSTCODE55	Y
35	20	MARRIED	Y	30	TERRACED	POSTCODE92	N
28	25	MARRIED	Y	70	DETACHED	POSTCODE87	N
36	17	MARRIED	Y	50	FLAT	POSTCODE18	N

Browse demo public data

- Select demo public GDSF
- Run select query
  - select * from demoPublicdata limit 50
- Query results appear
  - example public data

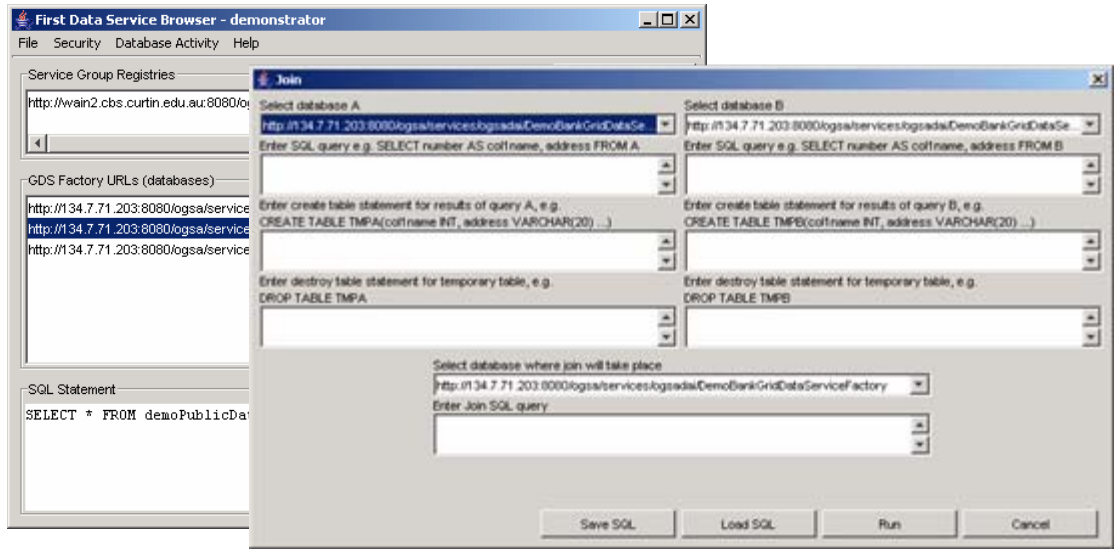




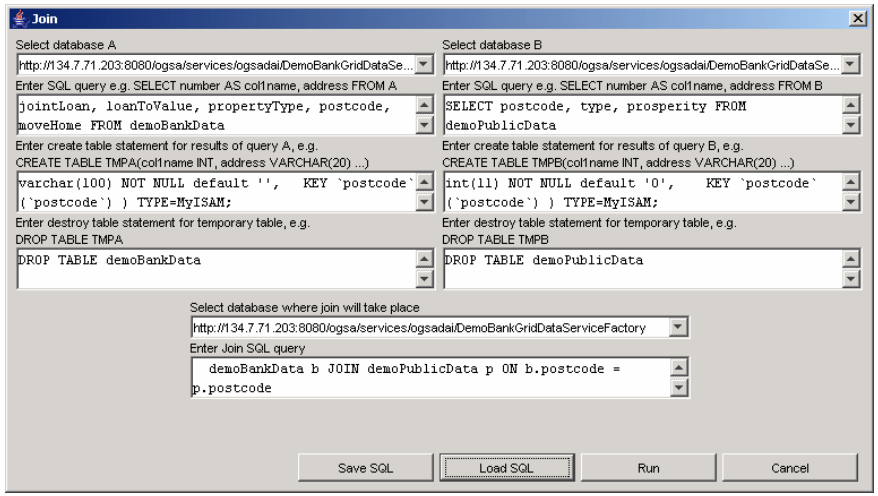
postcode	type	prosperity
POSTCODE0	CITY	6
POSTCODE1	CITY	4
POSTCODE2	TOWN	5
POSTCODE3	SUBURB	0
POSTCODE4	RURAL	5
POSTCODE5	CITY	1
POSTCODE6	SUBURB	1
POSTCODE7	RURAL	4
POSTCODE8	RURAL	5
POSTCODE9	TOWN	4
POSTCODE10	SUBURB	1
POSTCODE11	SUBURB	0
POSTCODE12	CITY	0
POSTCODE13	CITY	2
POSTCODE14	SUBURB	0
POSTCODE15	SUBURB	5
POSTCODE16	TOWN	1
POSTCODE17	TOWN	3
POSTCODE18	TOWN	4
POSTCODE19	SUBURB	6
POSTCODE20	TOWN	3
POSTCODE21	RURAL	1
POSTCODE22	SUBURB	4
POSTCODE23	SUBURB	1
POSTCODE24	TOWN	2
POSTCODE25	RURAL	5
POSTCODE26	CITY	5
POSTCODE27	SUBURB	0
POSTCODE28	CITY	4
POSTCODE29	CITY	3
POSTCODE30	SUBURB	0

Demo Data fusion

- Select Database Join activity

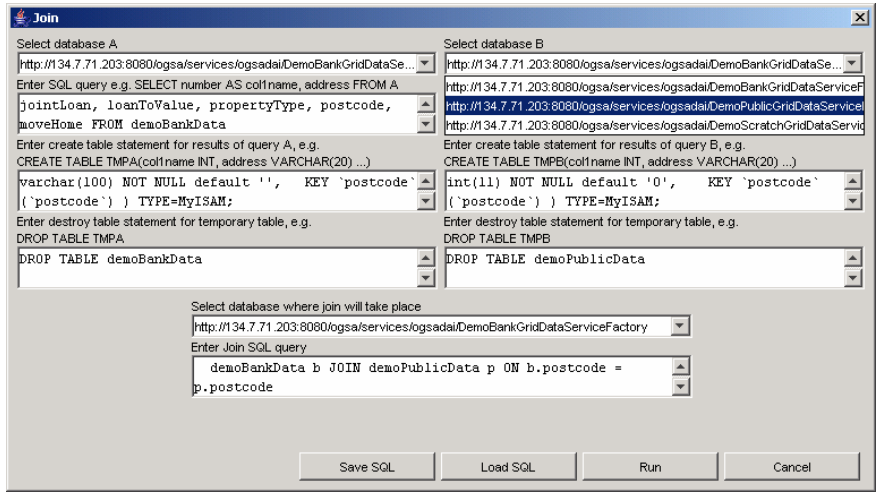


- Load SQL for data fusion pattern

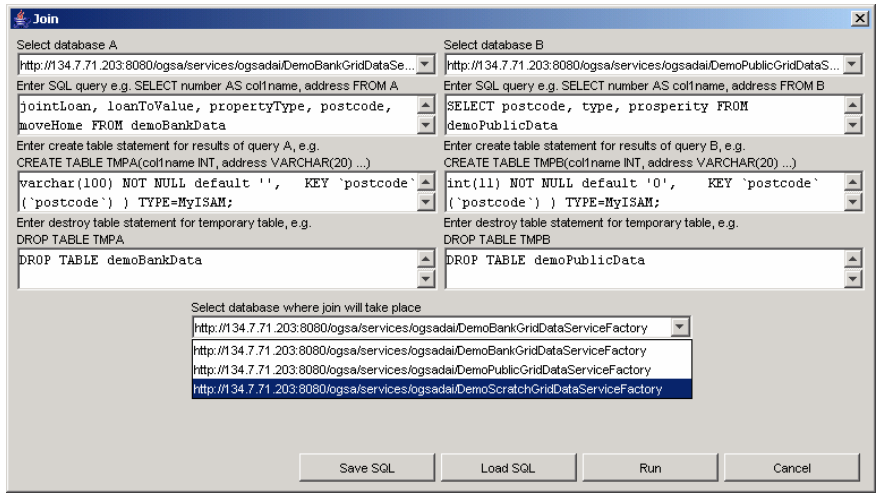


Demo Data fusion 2

- Configure join pattern
- Select source databases
- Join on postcode



- Set destination database



Data fusion results

Query Results

File

customerAge	loanDuration	maritalStatus	jointLoan	loanToValue	propertyType	regionType	regionProsp...	moveHome
49	11	SINGLE	N	70	DETACHED	CITY	6	Y
24	25	MARRIED	Y	90	FLAT	CITY	6	Y
24	25	MARRIED	Y	90	SEMI	CITY	6	Y
36	24	SINGLE	N	80	DETACHED	CITY	6	Y
48	12	MARRIED	Y	70	DETACHED	CITY	6	N
24	25	SINGLE	N	60	TERRACED	CITY	6	Y
21	25	SINGLE	N	90	FLAT	CITY	6	Y
35	25	SINGLE	Y	90	SEMI	CITY	6	Y
34	25	MARRIED	N	50	SEMI	CITY	6	N
41	19	MARRIED	Y	80	FLAT	CITY	6	N
32	10	MARRIED	N	100	TERRACED	CITY	6	N
34	25	SINGLE	Y	100	FLAT	CITY	6	Y
36	24	MARRIED	Y	60	DETACHED	CITY	6	N
19	25	SINGLE	Y	70	FLAT	CITY	6	Y
28	25	MARRIED	Y	80	TERRACED	CITY	6	N
29	14	MARRIED	Y	100	FLAT	CITY	6	N
30	25	SINGLE	N	90	FLAT	CITY	6	Y
51	10	MARRIED	Y	80	DETACHED	CITY	6	Y
26	25	SINGLE	N	30	TERRACED	CITY	6	N
34	25	MARRIED	Y	50	FLAT	CITY	6	N
25	25	MARRIED	Y	80	TERRACED	CITY	6	Y
35	24	MARRIED	Y	90	DETACHED	CITY	6	N
36	24	MARRIED	Y	50	DETACHED	CITY	6	N
18	25	MARRIED	Y	20	TERRACED	CITY	6	N
46	14	MARRIED	Y	70	DETACHED	CITY	6	N
29	18	MARRIED	Y	70	FLAT	CITY	6	N
36	24	SINGLE	N	90	FLAT	CITY	6	N
27	25	SINGLE	Y	30	SEMI	CITY	6	N
37	23	MARRIED	Y	60	SEMI	CITY	6	N
23	25	SINGLE	N	50	TERRACED	CITY	6	Y

Query Results

File

Minimize

postcode	type	prosperity
POSTCODE0	CITY	6
POSTCODE1	CITY	4
POSTCODE2	TOWN	5
POSTCODE3	SUBURB	0
POSTCODE4	RURAL	5

Query Results

File

customerAge	loanDuration	maritalStatus	jointLoan	loanToValue	propertyType	postcode	moveHome
19	25	SINGLE	Y	60	FLAT	POSTCODE44	Y
45	15	MARRIED	Y	70	DETACHED	POSTCODE94	N
28	25	MARRIED	Y	100	SEMI	POSTCODE77	Y
38	22	SINGLE	N	80	TERRACED	POSTCODE47	N
28	19	MARRIED	Y	50	DETACHED	POSTCODE58	N
47	13	MARRIED	Y	80	FLAT	POSTCODE34	Y
39	21	MARRIED	N	50	TERRACED	POSTCODE3	N
28	25	MARRIED	Y	60	DETACHED	POSTCODE77	N
33	14	MARRIED	Y	90	FLAT	POSTCODE55	Y
35	20	MARRIED	Y	30	TERRACED	POSTCODE92	N
28	25	MARRIED	Y	70	DETACHED	POSTCODE87	N
36	17	MARRIED	Y	50	FLAT	POSTCODE18	N
36	24	MARRIED	Y	90	TERRACED	POSTCODE37	N
38	22	MARRIED	N	50	DETACHED	POSTCODE18	N
37	23	MARRIED	Y	80	DETACHED	POSTCODE46	N
24	25	SINGLE	N	70	SEMI	POSTCODE5	Y
26	25	SINGLE	N	70	FLAT	POSTCODE37	Y
20	18	MARRIED	N	70	FLAT	POSTCODE92	Y
21	25	SINGLE	N	90	TERRACED	POSTCODE14	Y
27	25	MARRIED	Y	10	TERRACED	POSTCODE51	N
20	19	MARRIED	Y	60	DETACHED	POSTCODE12	Y
38	22	MARRIED	Y	80	TERRACED	POSTCODE69	N
24	25	SINGLE	N	80	TERRACED	POSTCODE21	Y
30	25	MARRIED	Y	80	SEMI	POSTCODE89	N
52	10	MARRIED	Y	80	DETACHED	POSTCODE82	N
49	11	SINGLE	N	70	DETACHED	POSTCODE0	Y
28	14	MARRIED	Y	60	DETACHED	POSTCODE17	Y
34	23	MARRIED	Y	60	DETACHED	POSTCODE76	N
25	25	SINGLE	N	80	DETACHED	POSTCODE45	Y
23	25	SINGLE	Y	60	SEMI	POSTCODE90	Y
22	25	SINGLE	Y	60	FLAT	POSTCODE82	Y

Barriers encountered

Barriers

- Trust
  - Dynamic, virtual organisation is simulated rather than created
  - Organisations understandably wary about installation of software and the access it provides
- Market
  - Not clear if data providers will publish data via web/grid service interfaces such as OGSA-DAI
- Security, Security, Security
  - Not mature enough
    - Bugs found in all major software used: Globus, OGSA-DAI and TOG
- Software
  - Not robust enough
    - OGSA-DAI V3.1 could not handle large results
    - Sys admin skills still necessary to maintain the grid

Lessons Learned

- Performing Data Integration:
  - TimeZone date problems
    - Dates are stored as a time so
      - 6:00am Dec 25th in Perth Australia is converted to
      - 10:00pm Dec 24th in Edinburgh, UK
      - If data is processed in the UK, the wrong date is used.
- Security issues:
  - As mentioned before Bugs in
    - Globus JavaCoG in GT3
    - OGSA-DAI could not switch security for Grid data transfers
    - TOG had no security option
  - All of these have been fixed
- Middleware not mature enough for commercial deployment
  - Not out-of-the box
  - Bug fixes were required
  - Scalability- difficulty with large results in OGSA-DAI V3.1
    - Fixed in OGSA-DAI V4.0

epcc

Conclusions

Conclusions

- Simulation explored the potential of a virtual organisation consisting of data providers and analytical scientists
- Grid-data fusion in global markets benefits from perceived strengths of the Grid in scope and (global) scale
- For this application, grid technologies not mature enough to support the operation of a dynamic, virtual organisation
  - Do not provide necessary security and robustness to instill trust
  - Still needs to establish a business benefit that outweighs the cost of addressing the risks(?)
- Project contacts
  - http://www.epcc.ed.ac.uk/inwa
  - inwa@epcc.ed.ac.uk

epcc

Future Plans

Future Plans

- Include Chinese Academy of Sciences (CNIC) as node in the INWA grid infrastructure – ESRC/Sun funded
- Upgrade from OGSA-DAI R3.1 to R4.0
  - Addresses security and performance issues
- Investigate ODBC connections to OGSA-DAI data services
  - ODBC typically available in the data analysis software used in business and social science research
- …then we can start to explore the impact of Grid capabilities on innovation processes and hence the Grid's potential to support (virtual) industry clusters

Sun microsystems

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL