ConvertGrid: Grid Enabling Population Datasets

Keith Cole
Keith.Cole@manchester.ac.uk

National Centre for e-Social Science (NCeSS) & MIMAS
University of Manchester

Presentation Overview

- Data Grids and the e-Social Science vision
- The ConvertGrid pilot demonstrator project
  - An example of Grid enabling population datasets & existing web based services
  - Lessons learned
- Building the Social Science Data Grid
  - The next steps
- GEMS project

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

MIMAS

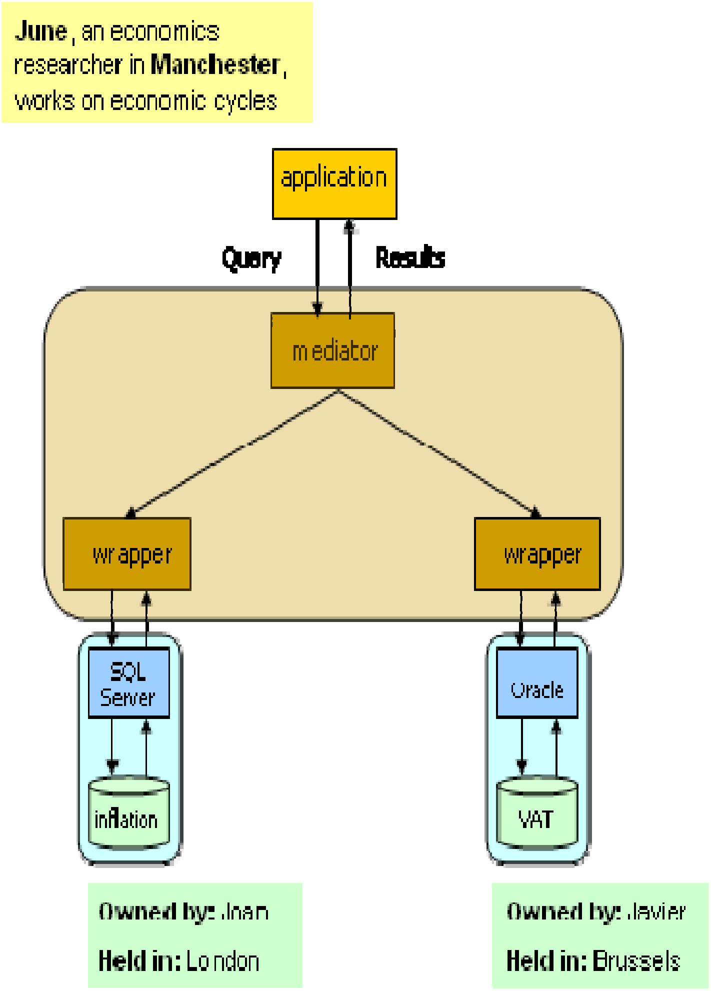
NCeSS National Centre for e-Social Science

What are the benefits of Data Grids for Social Science?

- Data Grids facilitate unimpeded use of distributed, heterogeneous, autonomous data resources.
  - Integrated view of the data resources that allow users to interact with them as if they constituted a single, global, integrated data resource.
- Grid enabling a dataset creates new opportunities for its use.
  - enables users to **integrate** it with other datasets
  - makes it possible to **analyse** the dataset using techniques that require the kind of computational power that it is only feasible using the Grid (e.g. more complex models, more data points).
  - standardisation of procedures and mechanisms used to access and update the dataset, increase its **shareability**

The Social Science Data Grid Vision

- It involves placing the data resource (e.g. database) behind ‘wrapper’ middleware.
- Once wrapped, ‘mediator’ middleware’ can be employed for data access.
- Once a data resource is Grid-enabled, its availability can be easily advertised in registries.
- June’s application can now access data on inflation and VAT as if Joan’s and Javier’s data were hers and held in Manchester.
- Analysis can be re-run automatically when databases are updated.
- It all sounds so easy in theory! Now let’s see a real example!



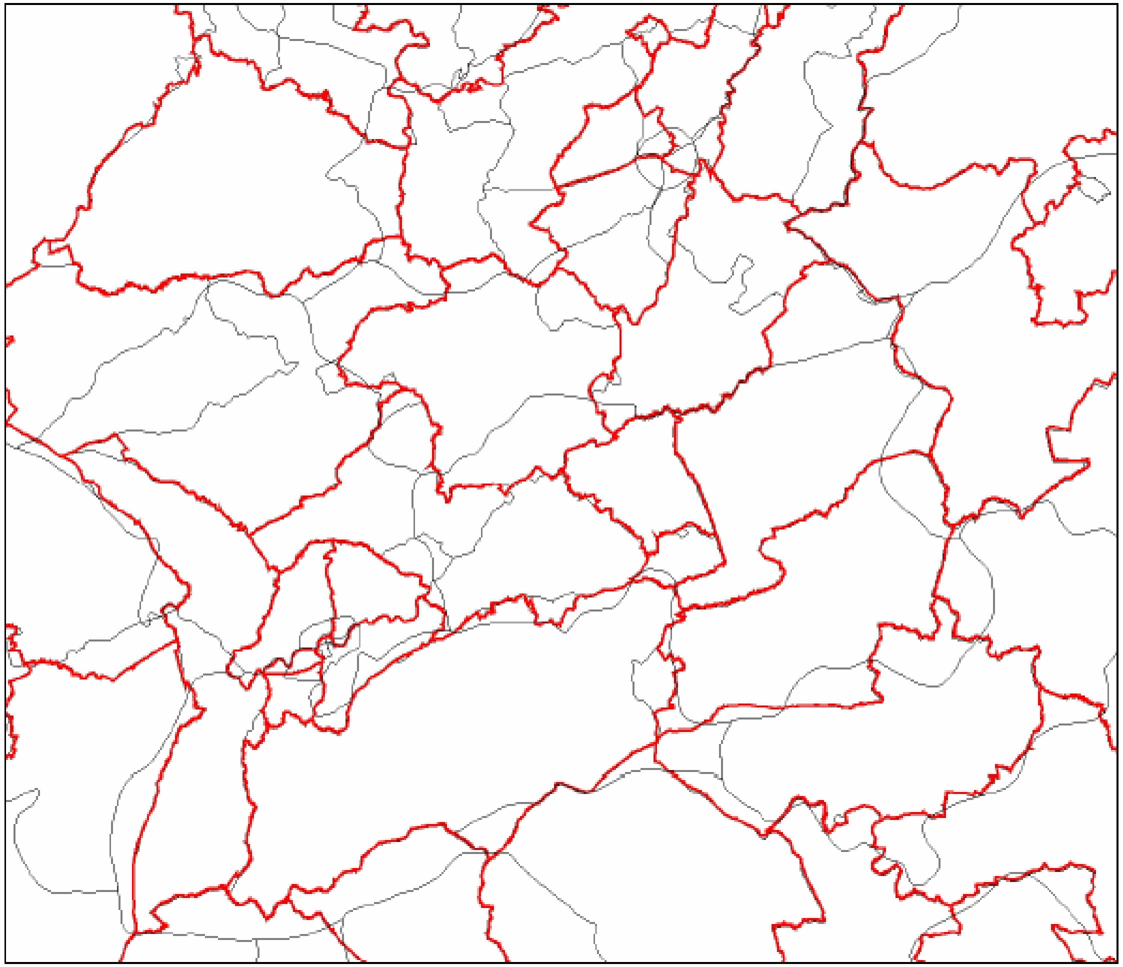
ConvertGrid – An e-Social Science Pilot Demonstrator Project

- Research context:
  - Research questions that require the combination of a data from multiple geo-referenced datasets which require users to perform the following generic tasks:
    - Extract data from a number of datasets using different interfaces
    - Convert each set of data to the desired target geography
    - Combine the converted sets into a single set of data
- ConvertGrid objectives:
  - To Grid enable existing socio-economic data sources;
  - Use Grid technologies to extend the functionality of an existing web based data service (i.e. Convert);
  - Demonstrate how Grid technologies can automate complex workflows;
  - Build a user interface to a Grid based service which is suitable for student/teaching use;

Different Target Geographies

1991 Wards

1991 Postcode Sectors



Source: Office for National Statistics

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

MIMAS

NCeSS National Centre for e-Social Science

ConvertGrid - Data Sources Used

- Data Sources
  - 1991 LBS/SAS (1991 Census geographies)
  - ONS Neighbourhood Statistics (1998 Ward & LADs)
  - Experian (2000 Postcode Sectors)
  - All Fields Postcode Directory (AFPD) (1999b)
- Selection criteria
  - Data on a range of themes (Health, Education and Crime Use Cases)
  - Different geographies and time points
  - AFPD derived conversion tables available for geographies via Convert

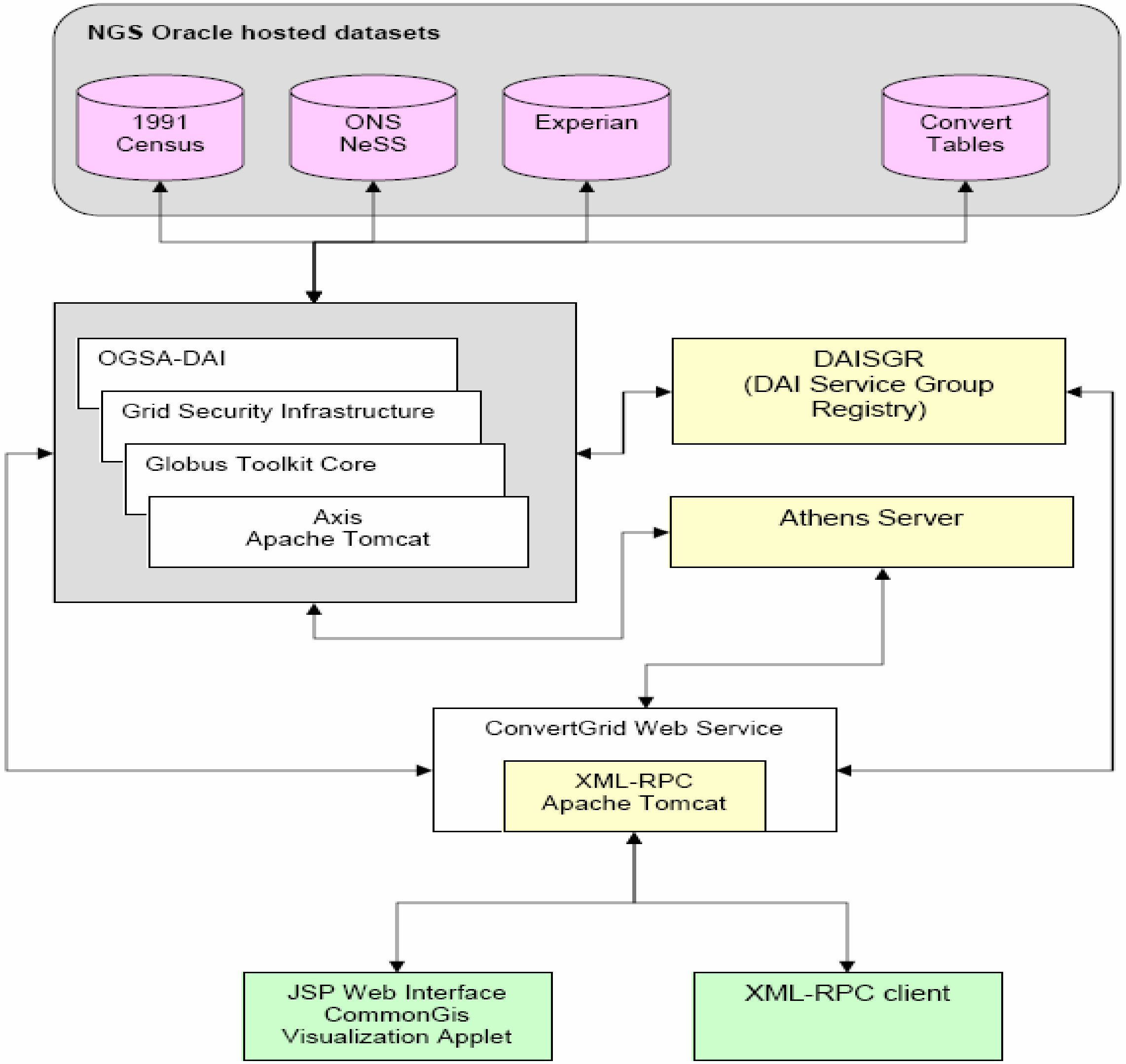
Example Use Case – Crime Theme

- Spatial correlation of recorded burglaries with house prices and other indicators of social wellbeing/deprivation.
- Study target geography –1998 LAD
- Datasets required:
  - 1991 Census
    - Total population (1991 ward)
    - Unemployment (1991 ward)
    - Overcrowding (1991 ward)
  - Neighbourhood Statistics 1998 data
    - Population estimates (1998 ward)
    - Recorded household burglaries (1998 LAD)
  - Experian1999 supply
    - Total population (1999 PCS)
    - Annual average house sale value (1999 PCS)
    - Population in MOSAIC Group A (1999 PCS)

Use Case – Health Theme

- Health researcher wishing to look for relations between incidence of coronary heart disease and other demographic factors.
- Study target geography –1998 Primary Care Group
- Datasets required:
  - 1991 Census
    - Total population (1991 ward)
    - Limiting Long Term Illness (1991 ward)
    - Unemployment (1991 ward)
    - Ethnicity (1991 ward)
  - Neighbourhood Statistics 1998 data
    - Population estimates (1998 ward)
    - Heart disease diagnosis episodes (1998 LAD)
  - Experian1999 supply
    - Total population (1999 PCS)
    - Population in MOSAIC Group A (1999 PCS)

ConvertGrid Architecture (*Techies only!*)



NGS Oracle hosted datasets

1991 Census

ONS NeSS

Experian

Convert Tables

OGSA-DAI

Grid Security Infrastructure

Globus Toolkit Core

Axis
Apache Tomcat

DAISGR
(DAI Service Group Registry)

Athens Server

ConvertGrid Web Service

XML-RPC
Apache Tomcat

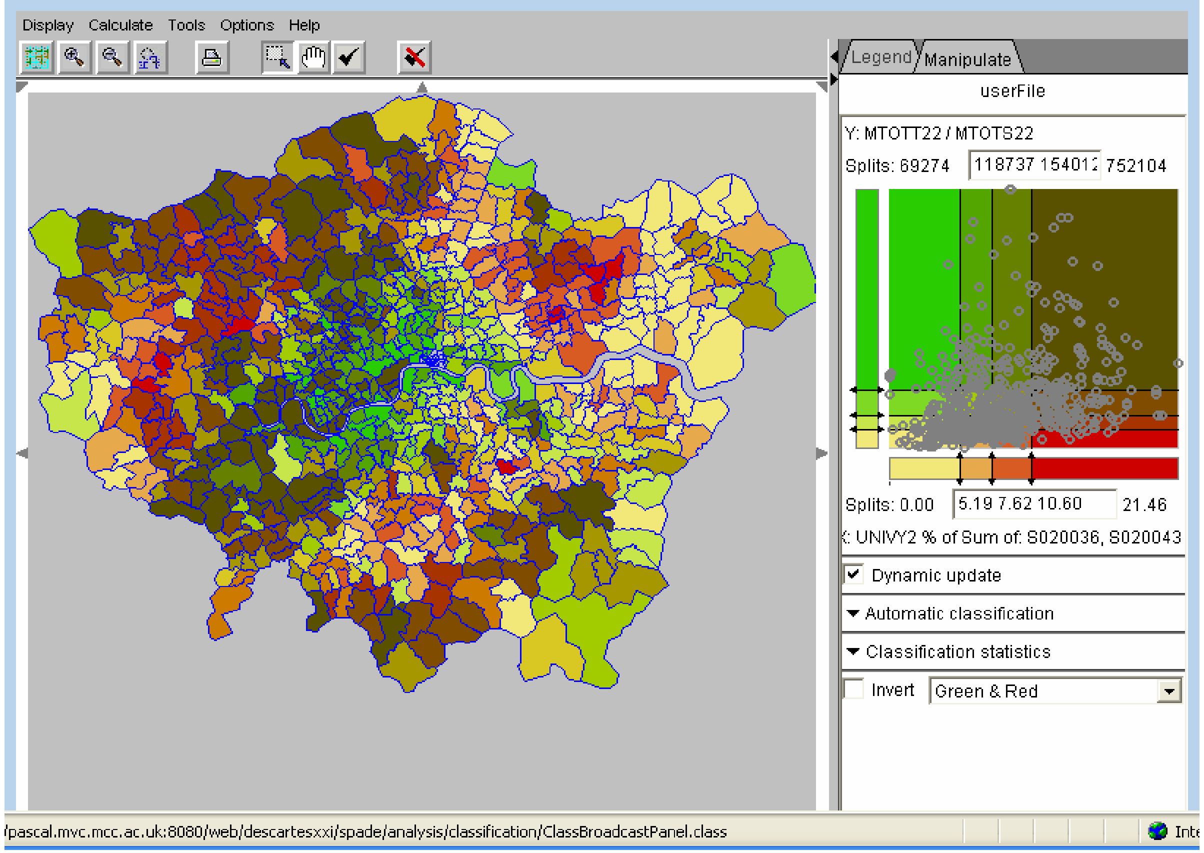
JSP Web Interface
CommonGis
Visualization Applet

XML-RPC client

ConvertGrid – Services Provided

- Converts data sources with different native geographies to a common Target Geography and outputs combined data as:
  - A data stream in CSV or XML format or
  - Transferred to a web based visualisation system
- Grid-enabled datasets (incl. AFPD)
  - Available to other Grid services
- Accessible to users via a ‘classic’ web based interface
  - Essential for demonstration purposes
  - Step by step guide developed
- Extensible system
  - Available to other applications via a web services interface
  - Easy to add other Grid-enabled datasets to the system

ConvertGrid – Data Visualisation Interface



userFile

Y: MTOTT22 / MTOTS22

Splits: 69274 118737 15401 752104

Splits: 0.00 5.19 7.62 10.60 21.46

: UNIVY2 % of Sum of: S020036, S020043

☑ Dynamic update

▾ Automatic classification

▾ Classification statistics

□ Invert Green & Red

/pascal.mvc.mcc.ac.uk:8080/web/descartesxxi/spade/analysis/classification/ClassBroadcastPanel.class

- Relationship between average house price sales (Experian) and percentage of 16-19 year olds entering university (Neighbourhood Statistics & Census aggregate statistics)

ConvertGrid – Issues and Challenges

- Establishment of a Grid infrastructure
  - Early adopter of the National Grid Service Data Node
  - Key Grid middleware still under rapid development
- Database migration problems
  - SQLServer to Oracle on the National Grid Service
  - Maintaining multiple databases resource intensive
- Data comparability issues a problem
  - Postcode formats
- Developing metadata registries
  - For resource discovery, data access and interpretation
- System performance, scalability and security
  - OGSA-DAI still relatively inefficient
  - Implementation of Grid security non-trivial

Grid Enabling Data - The Next Steps

- Establishing a social science data Grid is a key component of the wider e-Social Science strategy.
- Current social science data infrastructure (academic and non-academic) needs to be Grid enabled in a standards compliant and sustainable way.
- Data service infrastructures need to be able to support multiple forms of access.
- MIMAS is being funded by JISC to Grid enable the 2001 census aggregate statistics via OGSA-DAI on the NGS (GEMS project).
- The NCeSS Hub and Nodes will have a key role to play in addressing many of the key technical and methodological issues.
- Grid enabling the underlying databases may turn out to be the easy bit! Methodologies and intermediary applications/interfaces to facilitate data integration/analysis is much harder!

GEMS – Grid Enabling MIMAS Services

- Establishing production data grids to support e-Research
- Connecting the MS SQLServer databases holding the 2001 Census aggregate data directly to the Grid via the NGS
- Grid enabling the current data access system (Casweb)
- Maximise and build upon the ESRC/JISC investment in the establishment of an existing social science data infrastructure

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

MIMAS

NCeSS National Centre for e-Social Science

GEMS Functionality

- Transform query result into a variety of formats (CSV, HTML, etc...) by employing built-in or user uploaded XSL Transform scripts
- Upload query results to a Grid/FTP server
- View SQL generated by user interface for further integration into an OGSA-DAI client
- Redirect query results to an grid service/OGSA-DAI activity for further processing
- Bulk upload query results to a user specified OGSA-DAI enabled database
- Implement secure access management

Acknowledgements

- ConvertGrid Team @ Manchester
  - Jon Mclaren
  - Pascal Ekin
  - Linda Mason
  - Stephen Pickles
  - Justin Hayes
- NCeSS
  - Laura Bond
  - Alvaro Fernandes

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

MIMAS

NCeSS
National Centre for e-Social Science