

ConvertGrid – Data Grids for Social Science Research

Keith Cole

Keith.Cole@manchester.ac.uk

Presentation Overview

- Data Grids and the Social Sciences
 - Answers to some key questions
- The ConvertGrid Pilot Demonstrator Project
 - Objectives
 - Research context
 - A worked example
- Issues and Challenges
- Building the Social Science Data Grid – The Next Steps
- The GEMS Project

The Wider e-Science Vision

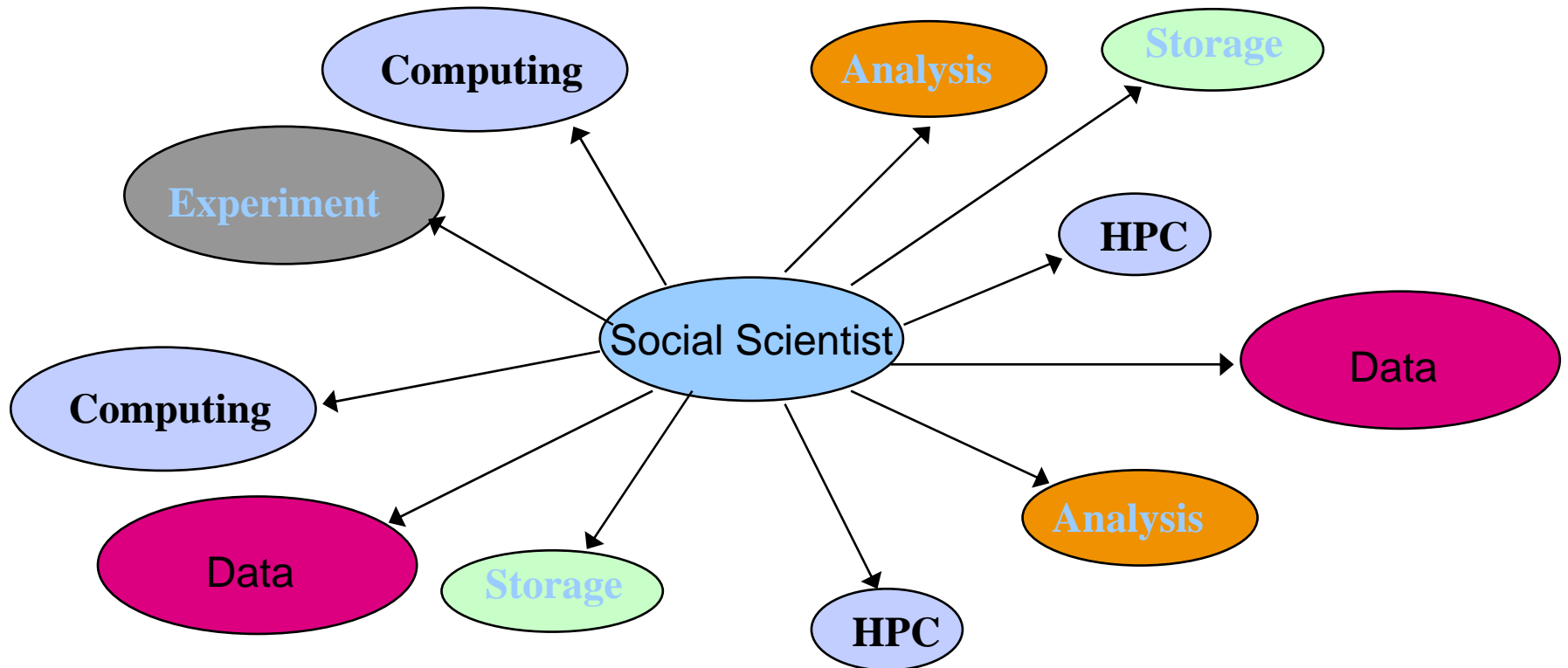
- Not “The web on steroids!”
- “e-Science is about global collaboration in key areas of science and the next generation of infrastructure that will enable it.” (John Taylor, former DG, Research Councils)
- That infrastructure is the **Grid**
- The Grid has the potential to facilitate a major step change in the way in which research is undertaken by:
 - Supporting novel forms of research
 - Enabling more complex forms of analysis
 - Automating complex workflows
 - Facilitating global collaboration to address global problems (e.g. global warming, genome research etc.)

Different Types of Grid

- Computational Grids
 - for scalable high performance computing resources (e.g. large scale simulations).
- Data Grids
 - for accessing, integrating and sharing heterogeneous datasets held in multiple locations.
- Access Grids
 - advanced video conferencing to facilitate collaboration between researchers nationally and internationally.
- Sensor Grids
 - for collecting real time data (e.g. traffic flows, electronic transactions).

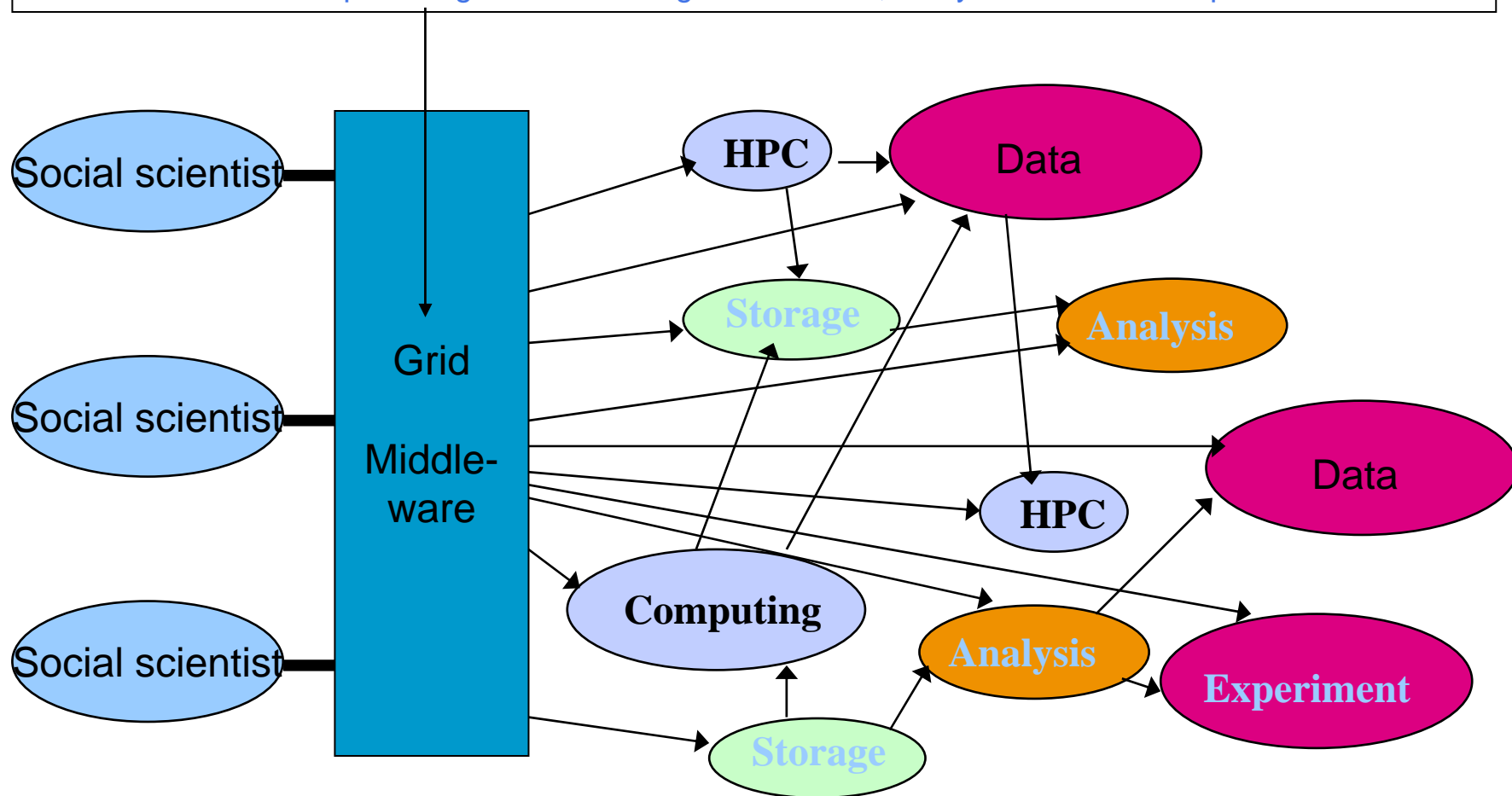
Social Science Research Infrastructure Today

- Many separate accesses, multiple architectures



Future Social Science Research Infrastructure?

Grid middleware manages the interactions between users and the heterogeneous and distributed resources on the Grid providing seamless integration of data, analytic tools and compute resources



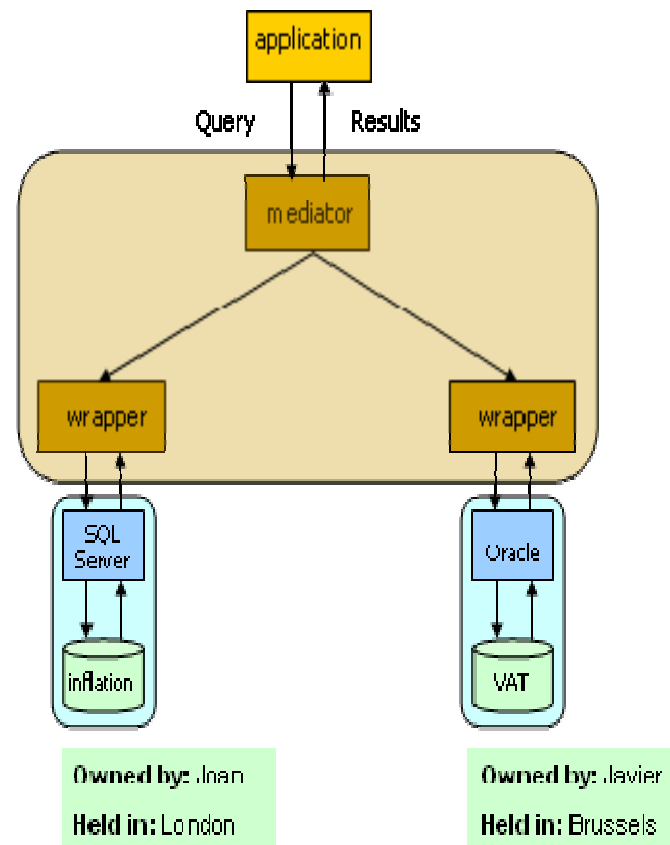
What are the Benefits of Data Grids for Social Science?

- Data Grids facilitate unimpeded and integrated use of distributed, heterogeneous, autonomous data resources.
- 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**
 - **Automated analyses** (i.e. analyses can be re-run automatically when databases are updated).

What Does Grid Enabling Data Entail?

- It involves placing the data resource (e.g. database) behind 'wrapper' middleware to provide a standard interface for data access (OGSA-DAI)
- Once wrapped, 'mediator' middleware can be employed for data access (OGSA-DQP)
- 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.
- For further information see <http://www.ncess.ac.uk/insight/tutorial/s/datagrids/>

June, an economics researcher in Manchester, works on economic cycles



ConvertGrid – Key Objectives

- Provide a practical demonstration of how the Grid can be used to facilitate data integration and overcome a major barrier to research use of multiple datasets;
- Demonstrate how to build a social science Data Grid by grid enabling a number of key geo-referenced socio-economic data sources;
- Use Grid technologies to extend the functionality of an existing web based data service (i.e. Convert) to exploit the existence of a Data Grid;
- Demonstrate how Grid technologies can automate complex workflows and enhance the capacity to address substantive social science research questions;
- Build a user interface to a Grid based service which is suitable for student/teaching use

ConvertGrid – The Research Context

- Many research questions require the combination of a data from multiple geo-referenced datasets
 - E.g. Linking post coded data to census geography
- The conversion of data relating to different geographies to a common target geography is
 - A complex time consuming task
 - Requires a range of data handling/processing skills
 - A major barrier to use!
- The data conversion process will require users to perform the following generic tasks:
 - Extract and download data in different formats from a number of databases using different interfaces
 - Convert each dataset to the desired target geography using geographical conversion tables
 - Combine the converted sets into a single dataset for analysis
- These generic tasks can be automated!

Different Source Geographies

- 1991 Wards
- 1991 Postcode Sectors



Source: Office for National Statistics

Existing Convert Service

- Developed as part of an ESRC funded project and transferred into service @ MIMAS
- Provides access to 225 UK-wide geography conversion tables between census, electoral, administrative, postal, health and statistical geographies derived from the All Fields Postcode Directory.
- Facility to convert a researcher's data from one set of geographical units to another (e.g. from postcode geography to health geography)
- The data conversion is improved by proportional allocation for those source units that do not fit within a single target unit.
- Extensible system - further conversion tables from any source can be incorporated

ConvertGrid - Data Sources Used

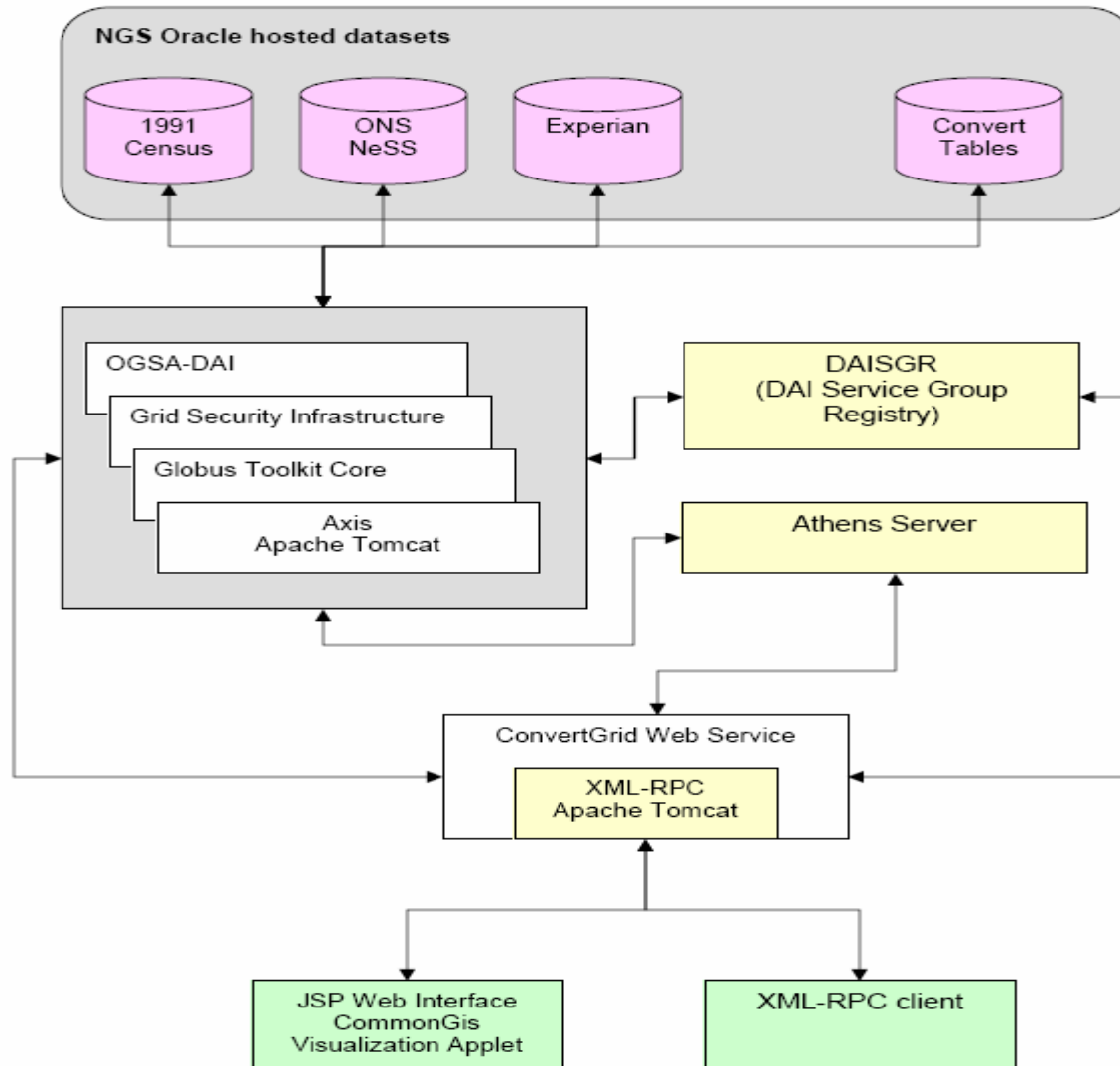
■ Data Sources

- 1991 Census Aggregate Statistics (1991 Census geographies)
- ONS Neighbourhood Statistics (1998 Ward & Districts)
- Experian (2000 Postcode Sectors)
- All Fields Postcode Directory (AFPD) (1999b)

■ Selection criteria

- Data on a range of themes to support Health, Education and Crime use cases.
- Different geographies and time points
- AFPD derived conversion tables available for geographies via Convert system

ConvertGrid Architecture (*Techies only!*)



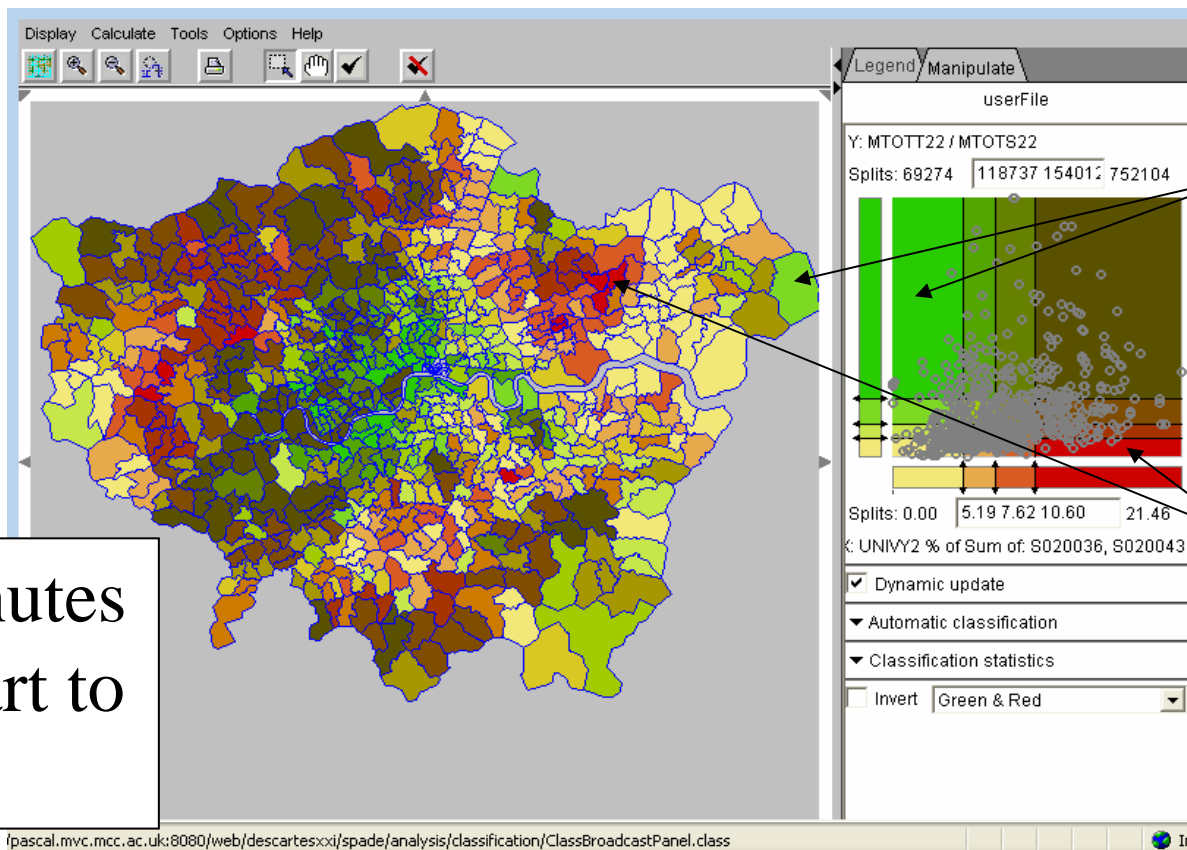
ConvertGrid – Services Provided

- 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 (Common GIS)
- Grid-enabled datasets (incl. AFPD)
 - Available to other Grid services via NGS
- Accessible to users via a ‘classic’ web based interface
 - 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 – A Worked Example

- What factors explain spatial variations in participation rates in higher education
- Study target geography –1991 Census Ward
- Data required:
 - 1991 Census
 - Total persons aged 16-17 & 18-19 (1991 Census Ward)
 - Neighbourhood Statistics
 - Number of applicants aged under 20 entering university (1998 Electoral Ward)
 - Experian
 - Average house price sales Quarter 2 2000 to Quarter 1 2001 (1999 Postcode Sectors)

ConvertGrid – Data Visualisation Interface



Ten minutes
from start to
finish

High
average
house price
sales but low
participation
rates

Low average
house price
sales but
high
participation
rates

- 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
 - Key Grid middleware immature and under rapid development
 - Performance, scalability and security issues
- Database migration problems
 - SQLServer to Oracle on the National Grid Service
 - Maintaining multiple databases resource intensive
- Data comparability issues a problem
 - Variable comparability
 - Postcode formats
- Developing metadata registries
 - For resource discovery, data access and interpretation

Building the Social Science Data Grid - The Next Steps (1)

- Establishing a production 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.
- Existing data service infrastructures need to be able to support multiple forms of access (i.e. single database approach)
- Still many technical problems to resolve
 - Mapping UK e-Science certificates to current and future access management protocols
 - Linking data and metadata

Building the Social Science Data Grid - The Next Steps (2)

- Managing user expectations is very important.
 - Data Grids do not overcome all barriers to research
 - Many demonstrators but few production services
- Complex, distributed workflows makes the development and deployment of services challenging and may require substantial software engineering!
- 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.
- Finally, MIMAS is being funded by JISC to Grid enable the 2001 Census aggregate statistics (GEMS project) as part of building a production Data Grid via the NGS.
 - 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)

GEMS Functionality

- Transform query result into a variety of formats (CSV, HTML, etc...) by employing built-in or user uploaded XSL Transform scripts
- Integration of table metadata into query results
- 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

Acknowledgements

- ConvertGrid & GEMS Teams @ Manchester
 - Pascal Ekin
 - Linda Mason
 - Stephen Pickles
 - Jon McLaren
 - Justin Hayes
 - Mat Ford (NGS)

- NCeSS
 - Laura Bond (NCeSS)
 - Alvaro Fernandes (Computer Science)