

Utilising a Grid Enabled Occupational Data Environment

GEODE – www.geode.stir.ac.uk

Paper presented to the XVIth ISA World Congress, Durban, 23-29 July 2006 – RC33 session 07, 'New Technologies and Data Collection in the Social Sciences'

| | |
|--|-------------------------------|
| Paul Lambert, Larry Tan, Ken Turner, & Vernon Gayle | University of Stirling |
| Ken Prandy | Cardiff University |
| Richard Sinnott | University of Glasgow |

'The Grid' and New Technologies of Data Collection

'The Grid' and 'eScience':

1. Online Coordination of electronic resources and collaborations
 - (Distributed computing)
 - Large scale
 - Collaborative
 - Heterogeneous
2. Standard protocols / information management systems

UK eSocial Science:

- 1) Investment in assessing / implementing technology
- 2) Computationally demanding data analysis
- 3) Qualitative and quantitative data collection technologies
- 4) ****Data sharing, processing and access****

GEODE: Survey records' occupational data

The importance of occupational micro-data(!)

Collecting occupational data

- 1) Initial occupational records (textual description)
- 2) Processing occupational records:

Text descriptions

- (1) Standardised Occupational Unit Group (OUGs)
- (2) Substantive occupational summary (e.g.,social class code)

Good practice:

- ✓ Preservation of original, OUG and substantive variables
- ✓ NSI's favour transparent occupational data coding (1) and translation systems (2)

Occupational data collection and processing

(1) Text records → OUG data

Currently:

Text coding software

(e.g. CASCOT)

Manual look-up

GEODE:

Linkage to existing resources

Further facilities possible but not planned (users typically have adequate resources)

(2) OUG data → summary indicators

Currently:

Numerous aggregate **occupational information resources**

Bespoke data programming requirements

GEODE:

Core provision: management and access of these data resources

Service to large volumes of users

Some illustrative occupational information resources

| | Index units | # distinct files (average size kb) | Updates? |
|--|---------------------|---------------------------------------|----------|
| CAMSIS, www.camsis.stir.ac.uk | Local OUG*(e.s.) | 200 (100) | y |
| CAMSIS value labels www.camsis.stir.ac.uk | Local OUG | 50 (50) | n |
| ISEI tools, home.fsw.vu.nl/~ganzeboom | Int. OUG | 20 (50) | y |
| E-Sec matrices www.iser.essex.ac.uk/esec | Int. OUG*(e.s.) | 20 (200) | n |
| Hakim gender seg codes (Hakim 1998) | Local OUG | 2 (paper) | n |

What's the problem?

| External user (micro-social data) | | | | Occ info (index file) (aggregate) | | | | User's output (micro-social data) | | | |
|--------------------------------------|-----|-----|---|--------------------------------------|------|------|------|--------------------------------------|-----|----|---|
| id | oug | sex | . | oug | CS-M | CS-F | EGP | id | oug | CS | . |
| 1 | 110 | 1 | . | 110 | 60 | 58 | I | 1 | 110 | 60 | . |
| 2 | 320 | 1 | . | 320 | 69 | 71 | II | 2 | 320 | 69 | . |
| 3 | 320 | 2 | . | 874 | 39 | 51 | VIIa | 3 | 320 | 71 | . |
| 4 | 874 | 1 | . | | | | | 4 | 874 | 39 | . |
| 5 | 874 | 2 | . | | | | | 5 | 874 | 51 | . |

*Indexed mainly by **Occupational Unit Group (OUG)**. But...*

- **Numerous alternative occupational data files** (time; country; format)
- **Alternative OUG schemes**; other index factors ('employment status')
- Inconsistent translations to social classifications – 'by file or by fiat'
- **Dynamic updates** to occupational data resources
- **Low uptake of existing occupational information resources**
- **Strict security** constraints on users' micro-social survey data

GEODE: Grid Enabled Occupational Data Environment

Strategy:

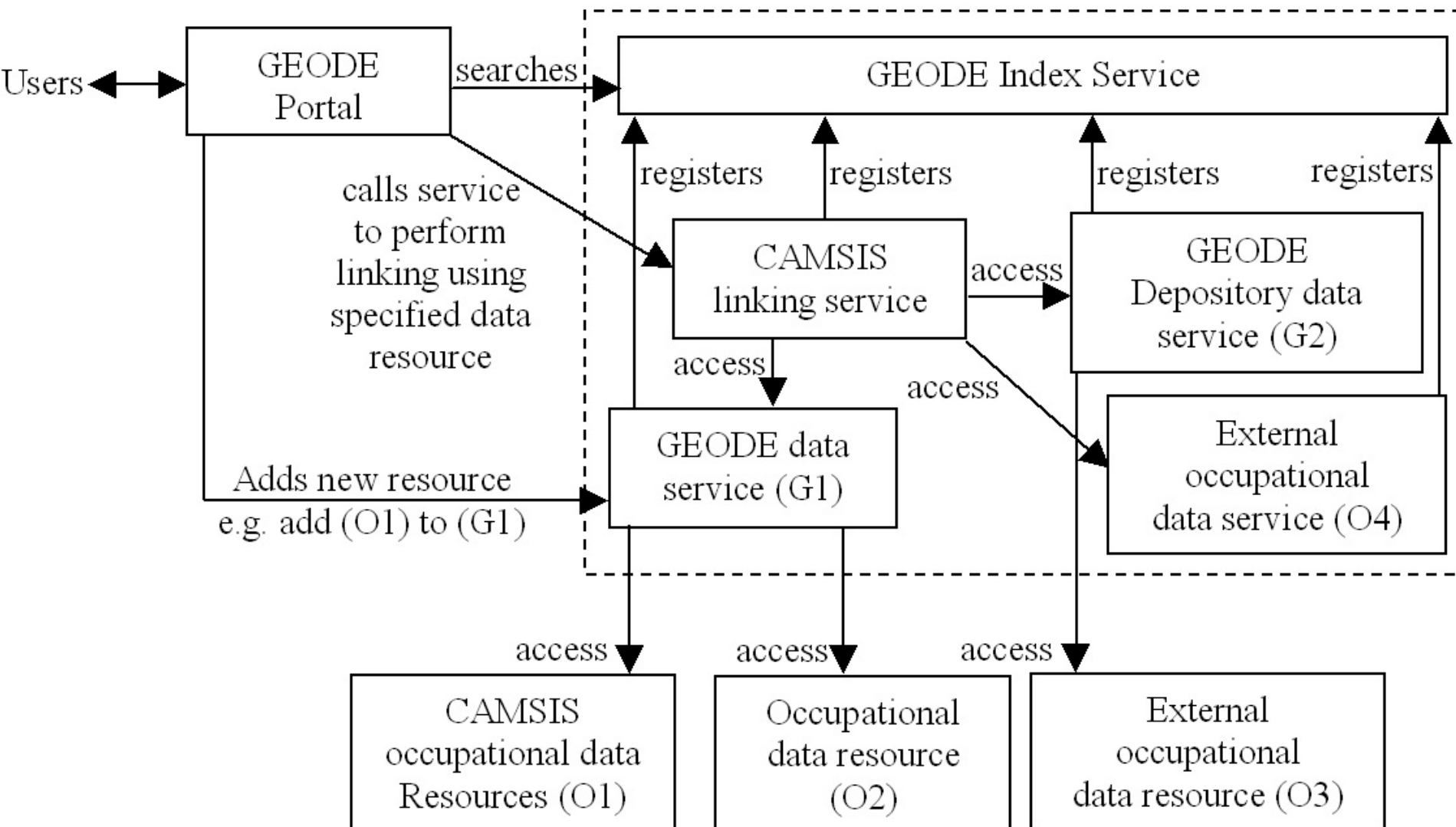
1) Occupational data index service (depository)

- i. Semantic data curation (DDI)
- ii. Data storage (OGSA-DAI)
- iii. Data indexing / access (OGSA-DAI)

2) User-friendly 'portal' access

- Entry to an international virtual organisation for data depositors and users (GridSphere, GT4, OGSA-DAI)
- Facilitate linking occupational information to users' datasets (OGSA-DAI) (initial focus on CAMSIS resources)

GEODE - architecture



Occupational information depository

1.1) Semantic curation of occupational information

- Establish a 'GEODE-M' meta-data subset (.xml)
 - Founded on Michigan Data Documentation Initiative
- Minimise curation requirements
- Web proforma entry
 - [via Portal using Gridsphere]

| | |
|--|--|
| <docDscr> <i>Release date</i> | <stdyDscr> <i><u>Country</u></i> <i><u>Time period</u></i> <i>Author</i> |
| <fileDscr> <i>Format</i> | <otherMat> <i>Missing data</i> <i>Data extensions</i> |
| <dataDscr> <varGrp> <var> <i><u>OUG variable</u></i> <i><u>Other identifier variables</u></i> <i><u>Output variables</u></i> | |

Occupational information depository

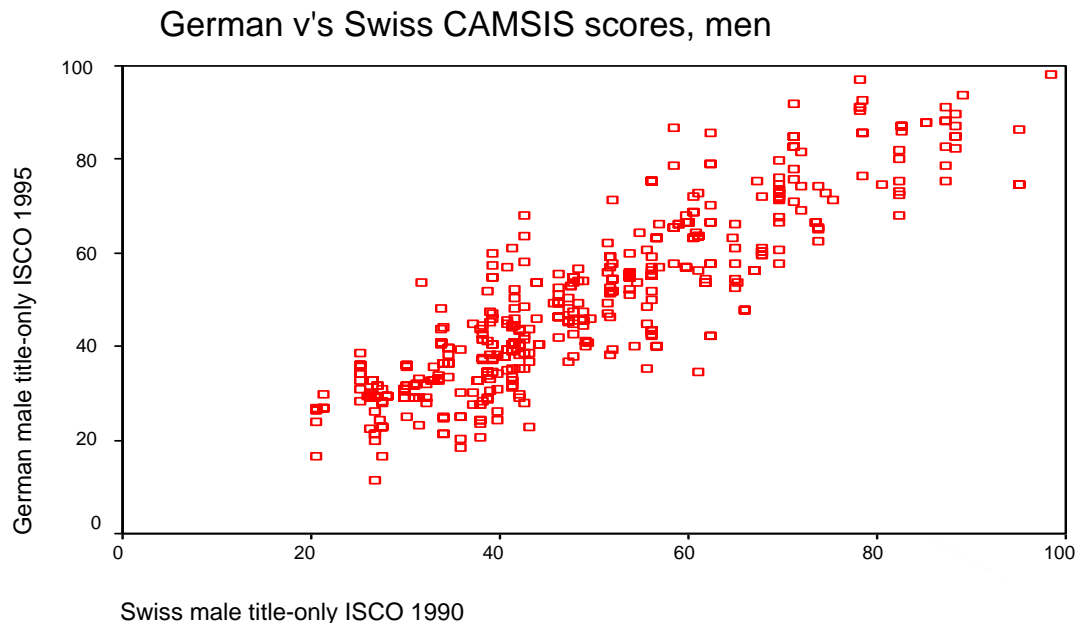
1.2) Storing occupational information resources

- GEODE-M documentation(2-stages)
- Storage: OGSA-DAI framework to link index files (dynamic)

Considerations:

- All data stored at GEODE v's Linkage to external data
- Proprietary software (*plain text / SPSS / STATA*)
- Rectangular index file?
- **plurality** of supply

⇒ **Universality or
Specificity?**



Occupational information depository

1.3) Virtual Organisation for Occupational Information Depository

- **MDS (via GT4)** to manage VO access to and distribution of occupational information resources
 - International virtual community
 - Dynamic data supply
 - OGSA-DAI efficient data indexing / searching / connecting
- *Grid: Create a community where members have abstract access to heterogeneous resources securely, and achieve wider collaboration*

2) Access to Occupational Data

2.1) File linkage mechanisms

Micro-social data (A) ↔ Occupational information resources (B)

- Multiple occupational variables on (A)
 - Strict security constraints on (A)
 - Inconsistent OUG formats on (A)
- Prototype linkages (e.g. CAMSIS) require full access to (A)
- Cater to limited access to (A):
- Investigate digital certification (X.509) to allow restricted data transfer A_[OUGs] + A_[context]
 - Requirements analysis
 - Minimal user certification process
 - Avoid application installation by users
 - Users' complex survey data (e.g. multiple occupational records)

GEODE portal access

2.2) Analytical queries

Process analytical tasks on aggregate occupational information resources

➤ Summary data

- Coverage searches
- Summary statistics

? Consider more complex analyses?

- CAMSIS derivations
- Involve interactive data management tasks
- *[cf. Nesstar / Data Web]*

Summary: GEODE services, www.geode.stir.ac.uk

- **Data collection service**
 - hinges upon curation of occupational information
 - User-friendly depository for occupational information resources
- **Data processing service**
 - User-friendly file matching facilities
 - Use of Grid to address file security concerns
- **Improved standards in occupational information utilisation**
- **Generalisability**
 - other information services, e.g., geographical; educational
- **eSocial Science**
 - Piloting of OGSA-DAI (with messy application)
 - Promotion of eScience facilities
 - Promising role with data construction process