

Development of a Grid Enabled Occupational Data Environment

GEODE – www.geode.stir.ac.uk

Paper presented to the Second International Conference on e-Social Science, Manchester, 28-30 June 2006

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

Development of a Grid Enabled Occupational Data Environment

1. Introduction: Occupational Information

Activities in two areas:

2. Occupational Information Depository

3. Access to occupational information

4. Conclusions and prospects

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

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

GEODE: Grid Enabled Occupational Data Environment

Objectives:

1) Operate as a portal

- Facilitate linking occupational information to users' datasets
 - (initial focus on CAMSIS occupational information resources)
- GEODE data resources – occupational information data curated as data service in Stirling, accessed by users via portal

2) Create an international Virtual Organization for occupational data community

- Sharing, indexing, & curating diverse occupational data
- Other analytical functions on occupational data?

GEODE – Building blocks

- **Globus Toolkit 4** (WSRF implementation)
 - To build grid application services
- **GridSphere 2.1.2** (portal framework – JSR 168)
- **OGSA-DAI** (data access grid middleware)
 - <http://www.ogsadai.org.uk/>
- **DDI** (social science metadata in XML)
 - <http://www.icpsr.umich.edu/DDI/>
- **Development environment:**
 - Jakarta Tomcat 5.x
 - Axis SOAP Engine
 - Java

2) Occupational Information Depository

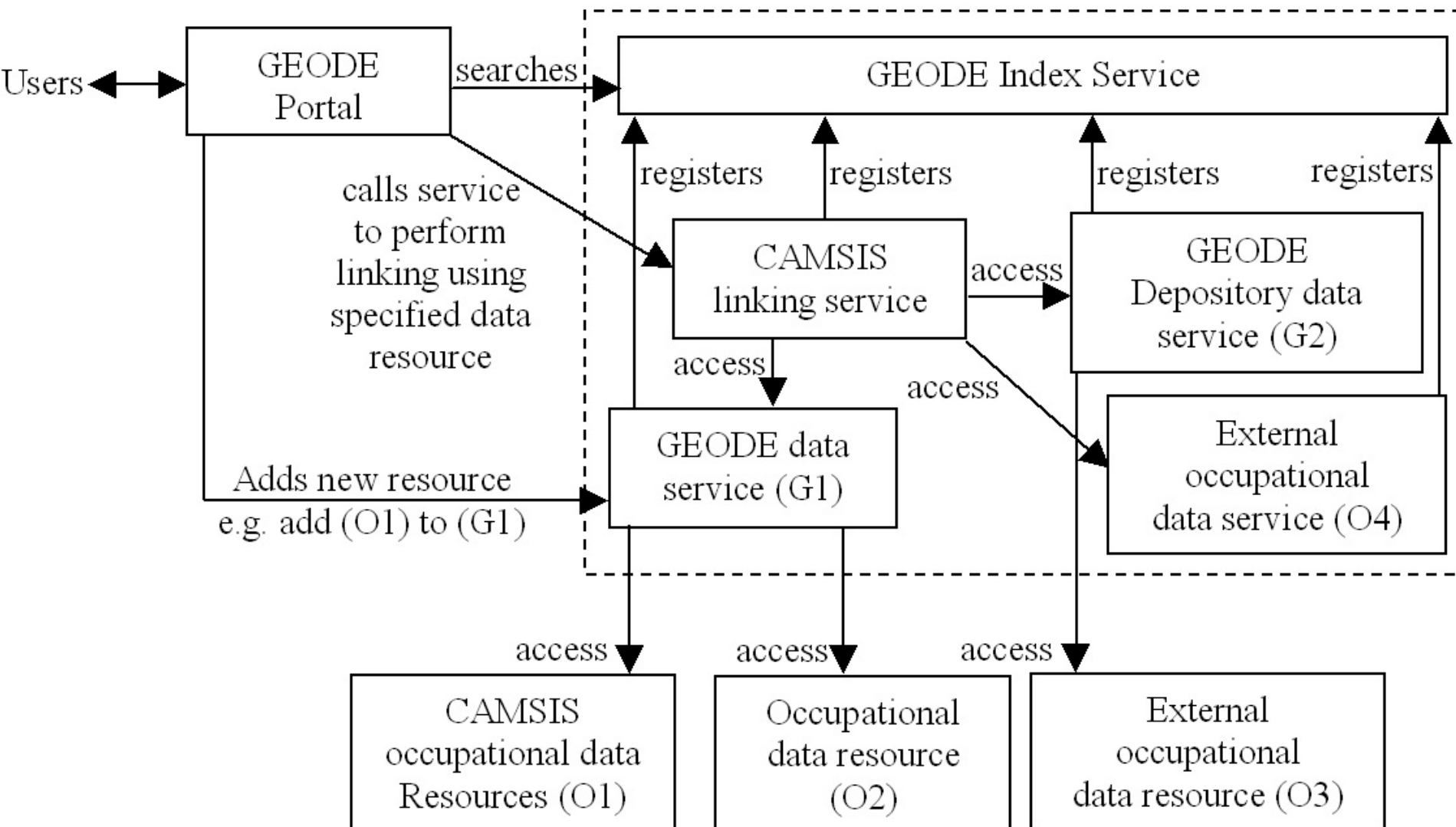
Grid as a system that... (e.g. Foster et al 2001):

- coordinates resources that are not subject to centralized control
- uses standard, open, general-purpose protocols and interfaces
- {delivers non-trivial qualities of service}

Use with occupational information depository:

- Create a community where members have abstract access to heterogeneous resources securely, and achieve wider collaboration

GEODE - architecture



GEODE: Occupational Information Depository

- Data Index Service – uses DDI and OGSA-DAI
- User Requirements / Evaluations
- Three elements:
 - 1) Semantic data curation
 - 2) Data storage
 - 3) Data indexing / access

Occupational information depository

2.1) Semantic curation of occupational information

- Establish a 'GEODE-M' meta-data subset (.xml)
 - Founded on Michigan Data Documentation Initiative
- Minimise curation requirements to suit occ. information resource providers (pilots)
- 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> <i><u>OUG variable</u></i> <i><u>Other identifier variables</u></i> <i><u>Output variables</u></i>	

Occupational information depository

2.2) Storing occupational information resources

Considerations:

- All data stored at GEODE v's Linkage to external data
- Proprietary software (*plain text* / *SPSS* / *STATA*)
- Rectangular index files v's other formats (e.g. pdf)
 - 'index file' format is easy and aids data storage / indexing
- Finite number of occ info. files / model of **plurality** of supply
- International community of data providers
- Negligible security restrictions (free online resources)

Strategy:

- 1) GEODE-M proforma, suits all formats, completed online
- 2) Translation to csv **index file**
- 3) Modify GEODE-M record for index file
 - (2) & (3) *performed automatically or manually*
- 4) Storage: OGSA-DAI framework to link index files

Occupational information depository

OGSA-DAI implementations on prototype service:

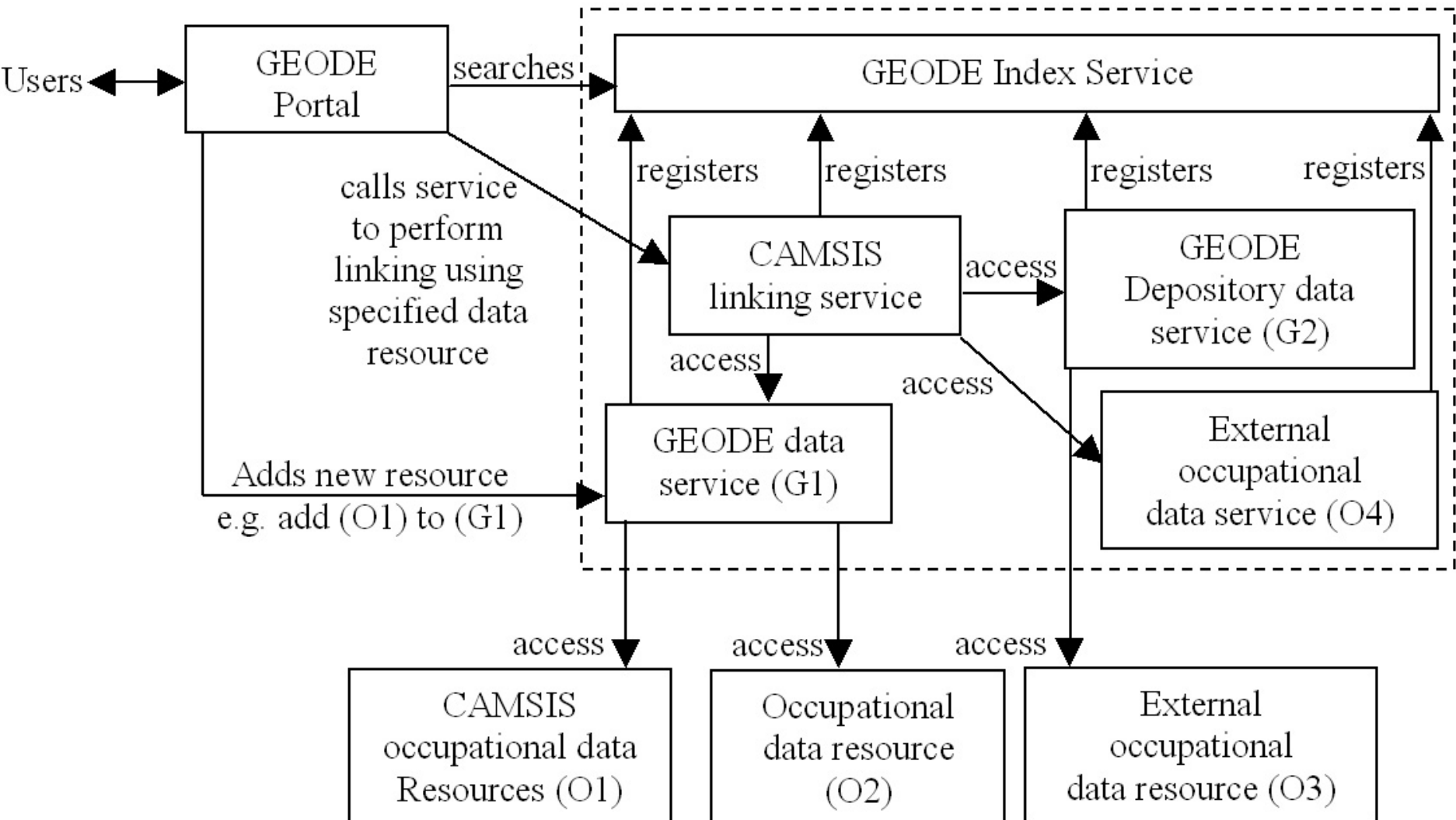
- Testing dynamic deployment of selected data resources (CAMSIS)
 - Registration with index service (pilot tests)
 - Searchable via portal service
- OGSA-DAI evaluations
 - **Foundations suited to collation of diverse occ data resources**
 - **Also facilitates data access functions (see 3)**
 - Accepts GEODE-curated resources; externally curated resources; and potential connections with other Grid data services
 - *{issues in support for alterative security levels to allow modification of initially deposited resources}*

Occupational information depository

2.3) 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

3) Access to Occupational Information



GEODE portal access

3.1) 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

3.2) 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]*

4) Conclusions and prospects

- **Occupational Information Depository**
 - OGSA-DAI implementations
 - Index-files annotated through 'GEODE-M'
 - Some ongoing manual support requirements
- **Portal framework**
 - Accessible GT4 / GSI structures...
- **Curation of occupational data**
 - Contribution: widely used international resources
 - Semantics: data annotation (DDI)
- **Generic data service**
 - Hinges on numeric OUG index [cf. [CASCOT](#)]
 - other application areas – e.g. Education, Geography

GEODE, eScience and eSocial Science

Some tentative comparisons...

	<i>Similar to</i>	<i>Diverges from</i>
GEODE-M metadata	DDI; Data Web; UKDA	[IDEAS]; [CAMSIS];
Data depository (OGSA-DAI)	ConvertGrid	[CASCOT]; [CAMSIS]; [ESDS]; Data Chronicles
Data matching service (OGSA-DAI + MDS)	[BRIDGES]	GEMEDA; Madiera
User engagement	Nesstar; ConvertGrid	CQeSS