Curating Occupational Information

GEODE – www.geode.stir.ac.uk

Grid Enabled Occupational Data Environment

Session 4 of GEODE Project workshop, 16th January 2007

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

Curating occupational information

- Assigning structure to ‘messy’ occupational information resources
  - Metadata on occupational information resources
  - Collating and defining occupational standard classifications
  - Lambert, P.S., Tan, K.L.T., Turner, K.J., Gayle, V., Sinnott, R.O. and Prandy, K. 2006. '**Data curation standards and the messy world of social science occupational information resources**' *Second International Digital Curation Conference*. Glasgow, www.dcc.ac.uk/events/dcc-2006/.
- Offering facilities for comparative occupational information
  - Lambert, P.S., Tan, K.L.T., Gayle, V., Prandy, K. and Turner, K.J. 2007 forthcoming. '**The importance of specificity in occupation-based social classifications**'. *International Journal of Sociology and Social Policy*.

Why is data on occupations 'messy'?

Messiness at both stages of the process:

1. **Collect & preserve 'source occupational data'**
2. **Summary / translation of source data**

- This model offers a 'scientific' approach
  - Published documentation (at both stages)
  - Replicable
  - Validation exercises
- But social researchers have been not been good at using it…
  - (Bechhofer 1969; Marsh 1986; Rose and Pevalin 2003)

{Stage 1 - Collecting Occupational Data and making a mess}

Example 1: BHPS			
Occ description	Employment status	SOC-2000	EMPST
Miner (coal)	Employee	8122	7
Police officer (Serg.)	Supervisor	3312	6
Electrical engineer	Employee	2123	7
Retail dealer (cars)	Self-employed w/e	1234	2

Example 2: European Social Survey, parent's data		
Occ description	SOC-2000	EMPST
Miner	?8122	?6/7
Police officer	?3312	?6/7
Engineer	??	??
Self employed businessman	??	?1/2

{Stage 1 - Collecting occupational data – summary}

- All methods lead eventually to coding to an occupational index scheme:
  - Occupational Unit Groups
  - Standardised Industrial Classifications
  - Standardised employment status classifications
  - *Somewhat less standardised occupational schemes*
  - *Not really at all standardised occupational index schemes*
- Occupational index schemes are the point of departure for GEODE

Stage 2 – using Occupational Information

Occupational information resources
are used to interpret occupational records

- Messy because:
  - Large volume of occupational information resources
  - Limited coordination between resources
  - Inconsistencies in access and exploitation processes

Occupational information resources

- **Large volumes** of occupational information resources
  - Coverage across countries and time periods
  - Different research fields / topics
  - Dynamic: updates to occupational information resources
    - Internet based distributions lead to duplication and expansion, e.g. ISEI - ISCO translation files at:
      - PISA webpages (Ganzeboom)
      - IDEAS/Repec webpagees (Hendrickx)
      - CAMSIS occupational data webpage

Some maths:

- 100+ alternative index schemes (OUGs; others)

X

- 500+ alternative output measures (class schemes, etc)

Occupational information resources

Limited coordination

- Varying metadata practices
  - Coordinated structure, e.g. ISEI at IDEAS/Repec [rare]
  - Natural language, e.g. CAMSIS [common]
  - No documentation
- Varying data file formats
  - SPSS, Stata, Plain text
- One-way distribution
  - Internet download; text publications
- Gaps between NSI's and academic researchers
  - NSI's make regular changes to favoured resources

Occupational information resources

Limited coordination (ctd)

- Varying translation rules
  - One file for all occupations ('universal')
  - Multiple files for different contexts ('specific')
  - Different occupational index requirements

ISEI	CAMSIS	EGP	Wright
{status scale}	*{stratification scale}*	*{class scheme}*	*{class scheme}*
Occ title	*Occ title; e.s.; gender*	*Occ title; e.s.*	*Occ conditions*

Occupational information resources

Inconsistencies in access / exploitation

- Occupational Unit Group schemes' variants
  - Decennial updates / International variations
  - Localised adaptations [e.g. HESA] / Survey variations [e.g. GHS]
  - Numeric or string format preservation
  - Hierarchical organisations
    - E.g. ISCO-88
    - 1234 ⊃ 123 ⊃ 12 ⊃ 1
    - 110 = 0110 ⊃ 11 ⊃ 1 ⊃ 0
- Focus for application of occupational data
    - Individual level measures
    - Household / career contexts

Returning to the occupational research model

Two stage process:

1. Collection & preservation of 'source occupational data'
2. Summary / translation of source data via occupational information resources

- Critically, stage (2) places responsibility for reviewing and treating occupational information resources with individual social scientists
- GEODE – alternative facility for managing stage (2)

Metadata - Occupational information information

How to facilitate searching, registering, accessing OIRs?

- **Establish a 'GEODE-M' meta-data subset** (.xml)
  - Founded on Michigan Data Documentation Initiative

- Semantic curation of occupational information
- XML convenient engagement with OGSA-DAI, Gridsphere, JAVA

<docDscr> *Release date*	**<stdyDscr>** *Country* *Time period* *Author*
<fileDscr> *Format*	**<otherMat>** *Missing data* *Data extensions*
<dataDscr> <varGrp><var> *<concept> to differentiate index and output variable groups* *<stdCatgry> to reference variable defintions*	

Example issues

- <StdCatgry> [Variant implementations <-> indexed translation files]
- <context> [cross-country resources]
- <producer> role="formatting" [caters to multiple author roles]
- <fileDscr id="dkcherisco88.sav"> [caters to multiple files]
- <abstract>

ISEI	CAMSIS	EGP	Wright
Occ title	*Occ title; e.s.; gender*	*Occ title; e.s.*	*Occ conditions*
***<stdCatgry>** (from www.geode.stir.ac.uk/ougs.html#)*			
ISCO88	*SOC90; ukempst; gdr*	*SOC90; ukempst*	*SIC92; SUPVIS; ..*
***<context>:** <nation abbr="..">* **<timePrd></timePrd>**			
10 [all]; **all**	*GB;* **1990-2000**	*GB;* **1950-2000**	*10 [all];* **1985-2000**

www.geode.stir.ac.uk/ougs.html

ISCO-88	International Standard Classification of Occupations
Description:	*ISCO-88 is a standardised occupational unit group scheme, published by the International Labour Office. It is designed to be implemented across countries and is recommended for international comparative research. Applicable from 1988 to present, it was designed as an update from earlier versions of ISCO developed in 1958 and 1968. The next revision to the ISCO scheme is anticipated for 2008.*
Structure:	*4-digit numerical index.* *Hierarchical structure incorporating 3-digit, 2-digit and 1-digit subgroups. Subgroups may be indexed both by truncation and by trailing zeros (e.g. ISCO-88 2445 'Psychologists' is part of minor group 244 'Social Science and related professions', which may also be indicated as minor group 2440 'Social Science and related professions').*
GEODE URI:	http://www.geode.stir.ac.uk/ougs.html#isco88
Examples:	1111 Legislators 2131 Computer systems designers and analysts 2142 Civil engineers 3413 Estate agents 5130 Personal care and related workers 5131 Child-care workers 8 PLANT AND MACHINE OPERATORS AND ASSEMBLERS 83 DRIVERS AND MOBILE-PLANT OPERATORS 832 MOTOR-VEHICLE DRIVERS 8322 Car, taxi and van drivers 8323 Bus and tram drivers
Internet links:	http://www.ilo.org/public/english/bureau/stat/class/isco.htm
References:	ILO (1990) *ISCO-88 : International Standard Classification of Occupations*, New York: International Labour Office.
Author of this note:	Paul Lambert, 22 August 2006

TOP

ISCO-88(COM)	International Standard Classification of Occupations, European Union Variant
Description:	*ISCO-88(COM) is a slight variant on ISCO-88. It is an standardised occupational unit group scheme, based upon the ISCO-88 scheme published by the International Labour Office, but with selected additional categories to allow integration with European Union statistical databases.*
Structure:	*As ISCO-88.*
GEODE URI:	http://www.geode.stir.ac.uk/ougs.html#isco88com
Examples:	1111 Legislators

Management of GEODE-M curation

Metadata considerations

- 'GEODE-M' as {flexible} recommended components of DDI
- GEODE-M templates
  - webpages at GEODE
  - Other facilities?

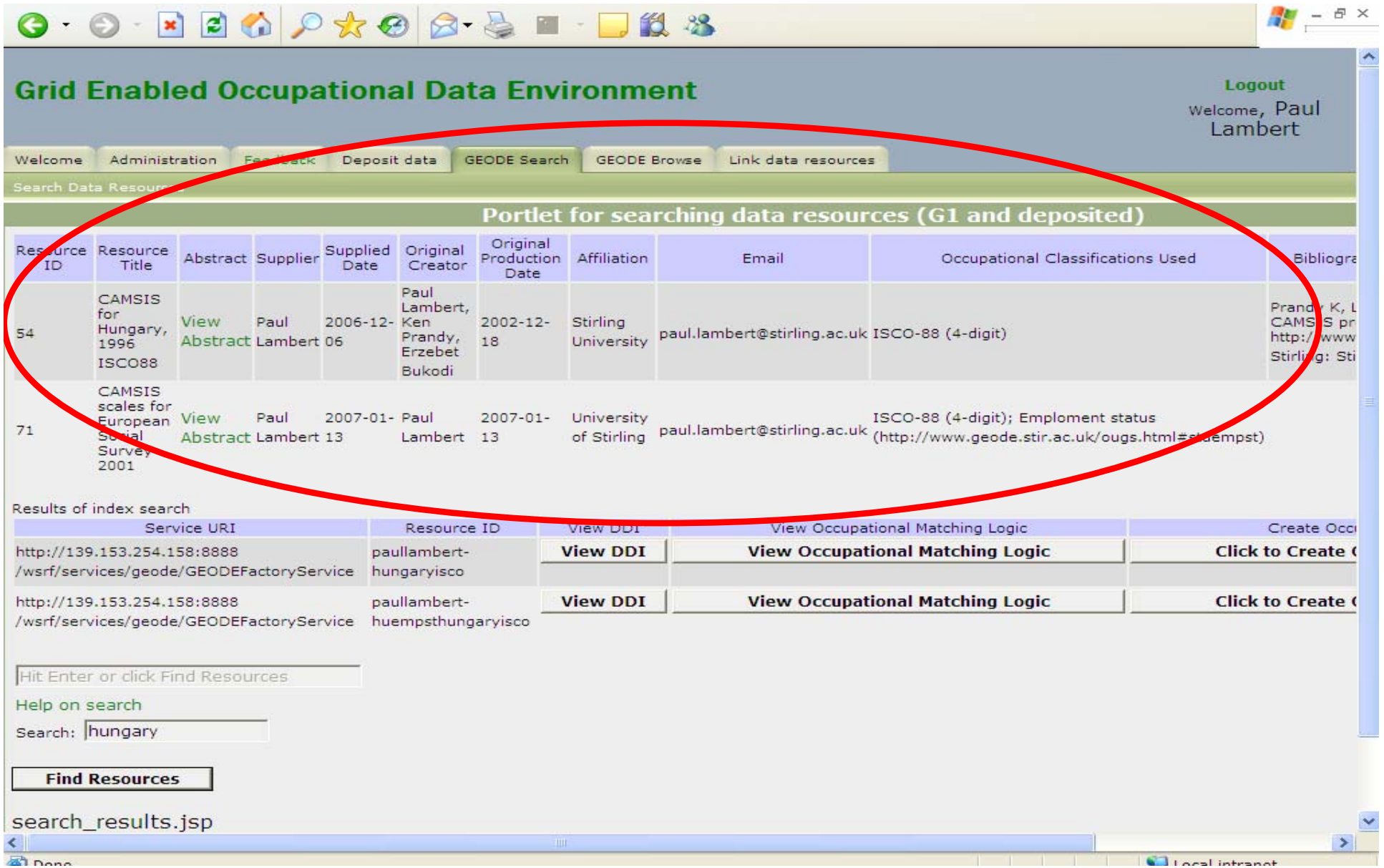
Data considerations:

- Stored at GEODE v's Linkage to external data

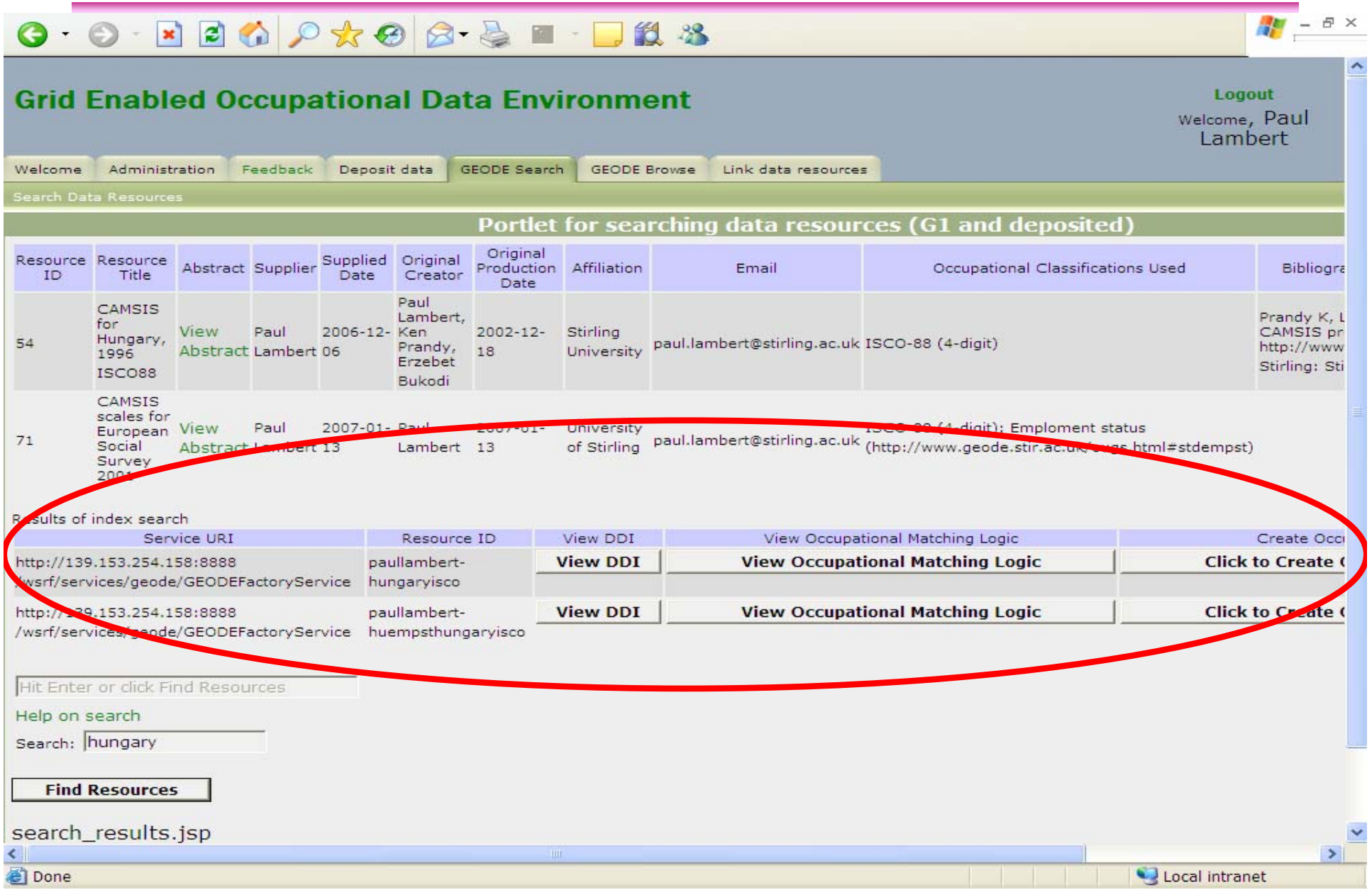
At present:

- Stage 1 – automated curation (allows external linkage, any file format)
- Stage 2 – extended manual curation (requires GEODE server copy of data, translation to plain text rectangular format
- Premised upon small commitment from depositors & GEODE

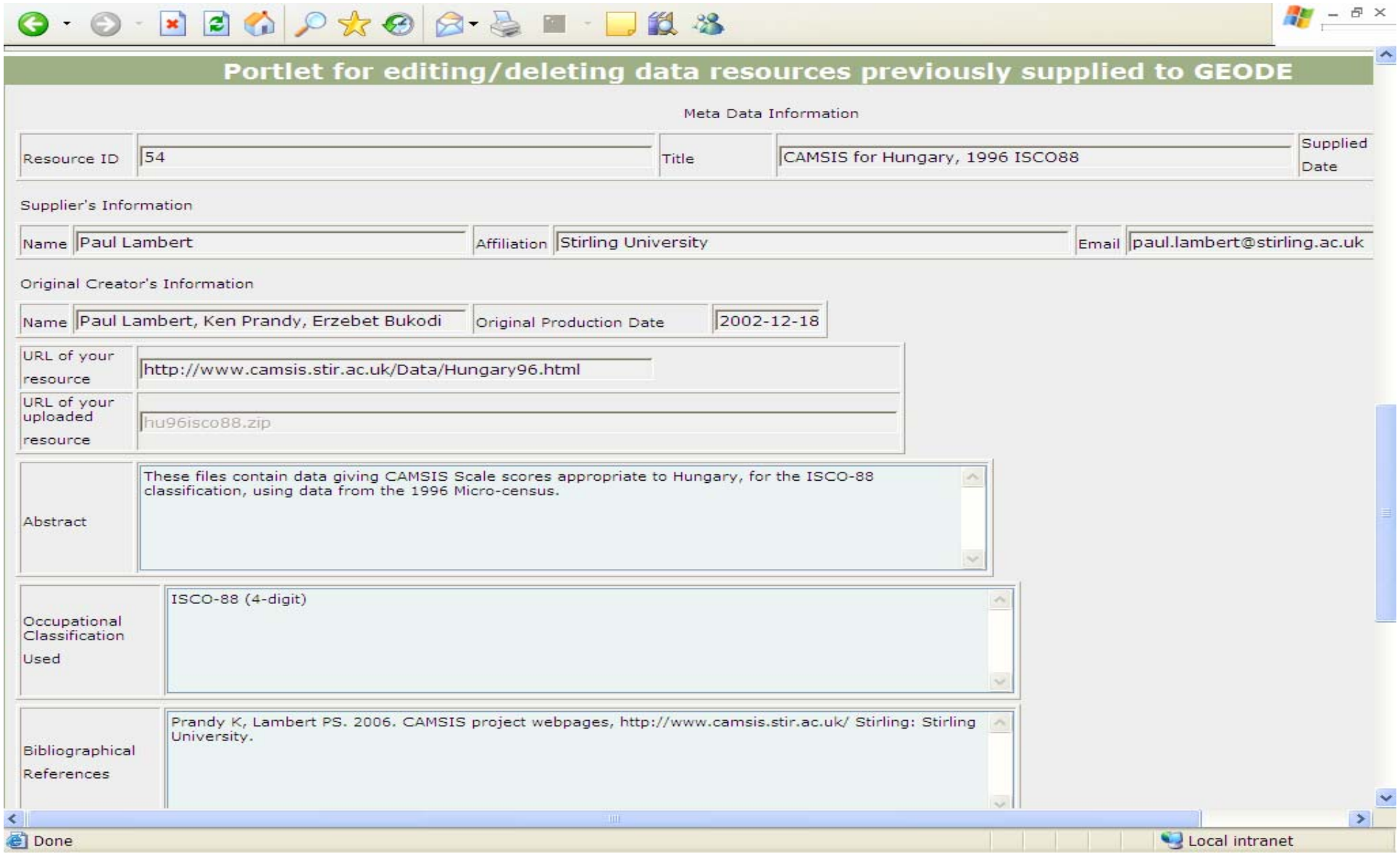
Searching – uncurated resources



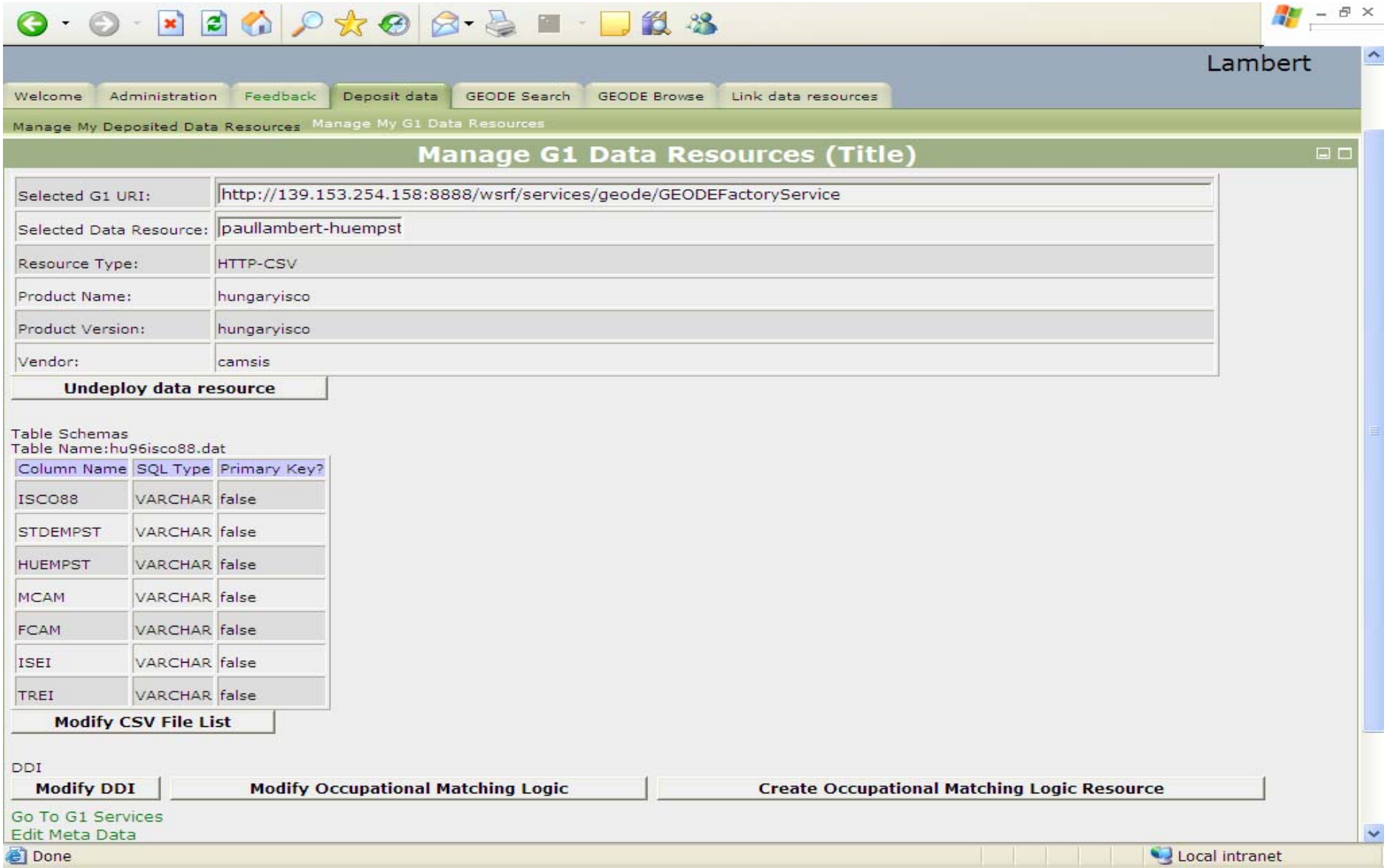
Searching – curated resources



Managing and modifying ‘uncurated’ resources



Managing and modifying ‘G1’ resources



Summary – assigning a structure to occupational information resources

- **Metadata**
  - xml format
  - DDI standard
  - 2-stage curation process

2) Comparative occupational information

GEODE Occupational Information Depository

- Collecting large volumes of OIRs from across countries, time periods
- Facilitation VO communication between occuaptional information resources

➢ Opportunity for evaluations of comparative occupational research

Universality and Specificity in social classifications

"Occupations are ranked in the same order in most nations and over time. ..Hout referred to the pattern of invariance as the "Treiman constant". ..the Treiman constant may be the only universal sociologists have discovered." (Hout and DiPrete, 2006:2-3)

"the idea of indexing a person's origin and destination by occupation is weakened if the *meaning* of being, say, a manual worker is not the same at origin and destination. Historical comparisons become unreliable" (Payne, 1992: 220, cited in Bottero, 2005:65)

Arguments for specificity

Theoretical

- Theories of change (over time, countries, gender)
- Theories of the minutae of occuaptional differences
- Widening scope of social science research
- more countries, time periods
- More micro-data resources

Empirical

- small increments to specific approaches
- broad equivalence across contexts

Universality

Comparative occupational research methods remain trenchantly universalist in principle:

- Forcing equivalent data collection / treatment across contexts
- *'The categories are different and it's not comparable'*

Why?

- **Substantial pragmatic hurdles to any other approach**

- *E.g. Cross National Equivalence File model*
  - **Model 1 (universal ISEI)**
    - CNEF data plus 1 file download; Approx 1.5k lines in Stata..
    - Approx 6 hours development
  - **Model 2 (specific - CAMSIS)**
    - CNEF data, plus original BHPS, PSID and GSOEP, plus 6 further file downloads; Approx 3k lines in Stata..
    - Approx 40 hours development / estimation

Universality v's Specificity

- *Limits of universality…*
  - Loss of the technological excuse…?
  - Sustainability of specific approaches
  - Need to engage with specific expectations
  - Contextuality of importance of specificity…

- GEODE contribution:
  - **Offers opportunity for specific approaches**
  - Potential generalisability for comparative research– education; geography

Conclusions

- Occupational data curation and the Grid
  - Grid facilitates management / access of occupational records via xml formats (OGSA-DAI)
  - Current models require moderate specialist input (manual curation)
  - Grid offers new level of service not previously available
    - Dynamic coordinated file storage
    - File matching [security]

- Comparative occupational analysis
  - New opportunities in occupational comparisons