A Data Curation Application Using DDI: The DAMES Data Curation Tool for Organising Specialist Social Science Data Resources

Simon Jones*, Guy Warner*, Paul Lambert**, Jesse Blum*

* Computing Science, ** Applied Social Science
University of Stirling, Stirling, Scotland, UK

IASSIST 2010
Session: Automated Curation Tools and Services for Metadata
Ithaca, NY       3 June 2010
DAMES: Background

- DAMES: Case studies, provision and support for data management in the social sciences
- A 3 year ESRC research “node”: Feb 2008-Feb 2011
  - Funded by the Digital Social Research Strategy group
- Driven by social science needs for support for advanced data management operations
- “In practice, social researchers often spend more time on data management than any other part of the research process” (Lambert)
- A ‘methodology’ of data management is relevant to ‘harmonisation’, ‘comparability’, ‘reproducibility’ in quantitative social science
DAMES: Themes

- Four social science themes
  - Grid Enabled Specialist Data Environments: occupations; education; ethnicity
  - Micro-simulation on social care data
  - Linking e-Health and social science databases
  - Training and interfaces for data management support

- Underlying computer science research themes
  - Metadata
  - Data curation
  - Data fusion
  - Workflow modelling
  - Data security
DAMES goals

- To build on previous work on the GEODE project
  - Grid Enabled Occupational Data Environment
- To build an integrated DAMES **Grid-based portal**:  
  - Combining specialist occupational, ethnicity, educational, health data resources: GE*DE(“GESDE”)  
  - Enabling a virtual community of soc sci researchers:  
  - To deposit and search heterogeneous data resources  
  - To access online services/‘tools’ that enable researchers to carry out repeatable and challenging data management techniques such as:    
    • fusion    • matching    • imputation    …
- **Facilitating access** is an important goal
DAMES scenarios

- Curation scenarios include:
  - Uploading occupational data to distribute across academic community
  - Recording data properties prior to undertaking data fusion involving a survey and an aggregate dataset

- Fusion scenarios include:
  - Linking a micro-social survey with aggregate occupational information (deterministic link)
  - Enhancing a survey dataset with ‘nearest match’ explanatory variables (probabilistic link)
Key role for metadata

- **Metadata records are absolutely core to the functioning of the portal infrastructure**
  - For adequate, searchable records for the heterogeneous resources (data tables, command files, notes and documentation)
  - To connect the resources and the data mgmt tools
  - To document the data sets resulting from application of the data mgmt tools: inputs, process, rationale,…

- **DAMES requirements:**
  - (Micro-)data based, very general, lifecycle oriented, Grid friendly
  - DDI 3
The metadata "cycle"
DAMES portal architecture overview

(Done: Security omitted)
Curation Tool prototype

The source data:

<table>
<thead>
<tr>
<th>Universal scales (derived from data from multiple countries)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>U1: Male and female, 1800-1938</td>
</tr>
<tr>
<td>U2: Male only, 1800-1938</td>
</tr>
<tr>
<td>U3: Female only, 1800-1938</td>
</tr>
<tr>
<td>E: Early period, 1800-c1890</td>
</tr>
<tr>
<td>L: Later period, c1890-1938</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
</tbody>
</table>
DAMES Curation Tool
Curate a data resource into the DAMES Infrastructure

Preliminaries
Metadata for this resource (e.g. DDI)
Have you begun curating this resource with us before, and/or do you have an existing structured metadata file in DDIII (Data Documentation Initiative) format for this resource?

Yes  No

This tool is now running in manual entry mode.

Short Description
Please provide a short description of this curation (40 characters max).

HIS-CAM scale for all countries (MS Excel file)
**DAMES Curation Tool**
Curate a data resource into the DAMES Infrastructure

### Preliminaries

**Metadata for this resource (e.g. DDI)**

Have you begun curating this resource with us before, and/or do you have an existing structured metadata file in DDI3 (Data Documentation Initiative) format for this resource?

- [ ] Yes
- [x] No

Please click "Browse" to select the DDI3 file. After selecting your file please click "Load" to fill in this form with the appropriate values.

- Browse
- Load

So as to provide you with appropriate options in later steps, please select the purpose of selecting this DDI file

- [ ] Edit existing
- [ ] New (to DAMES)
- [ ] Edit as new

### Short Description

Please provide a short description of this curation (40 characters max).

```
HIS-CAM scale for all countries [MS Excel file]
```
### DAMES Curation Tool
Curate a data resource into the DAMES infrastructure

#### Original Authors
All fields on this page are optional.

**Original Creation Date:** 19/10/2009

**Notes:** Name can be a person or group name. Multiple authors may be added (max 15).

<table>
<thead>
<tr>
<th>Author1</th>
<th>Author2</th>
<th>Author3</th>
<th>Author4</th>
<th>Author5</th>
<th>Author6</th>
<th>Author7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Paul Lambert</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization</td>
<td>University of Stirling</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organization webpage</td>
<td><a href="http://www.camsis.stir.ac.uk">www.camsis.stir.ac.uk</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact e-mail</td>
<td><a href="mailto:paul.lambert@stirling.ac.uk">paul.lambert@stirling.ac.uk</a></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DAMES Curation Tool
Curate a data resource into the DAMES Infrastructure

Occupational Classification - Country

Select country

Country List

<table>
<thead>
<tr>
<th>Country</th>
<th>EUds</th>
<th>Sweden</th>
<th>France</th>
<th>Germany</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ukraine</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>United Kingdom</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uruguay</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uzbekistan</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vanuatu</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venezuela</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vietnam</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yemen</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zambia</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zimbabwe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Most Common
Common Aggregations
Common Regions
Af - Bo
Br - Cu
Cy - Ge
Gh - It
Ja - Li
Lu - Mo
Na - Pa
Pe - Se
Si - Sw
Sy - Ug
Uk - Zi
None Defined
DAMES Curation Tool
Curate a data resource into the DAMES Infrastructure

Occupational Classification - Standards Used

Please provide details of any custom classifications used in the editor under “Custom”.

Classification List

- International
  - Australia
  - France
  - Germany
  - Ireland
  - Netherlands
  - Sweden
  - United States

Add classification
### Occupational Classification - Standards Used

Please provide details of any custom classifications used in the editor under "Custom".

- **HISCO (5-digit)**

**Historical International Standard Classification of Occupations**

| Description | HISCO is a standardised occupational unit group scheme designed to be most suitable for data from the periods 1600-1945. It is published by the International Institute for Social History. Its structure and definitions are loosely modelled upon ISCO-88, with some adjustments. |
| Structure | 5-digit numerical index. Hierarchical structure including component major groups (1-digit), submajor groups (2-digit) and minor groups (3-digit). |
| GEODE URI | http://www.geode.stir.ac.uk/ougs.html#hisco |
| Examples |  |
| Internet links | http://historyofwork.isg.nl/ |
| Author of this note | Paul Lambert, 21 September 2006 |
DAMES Curation Tool
Curate a data resource into the DAMES Infrastructure

<table>
<thead>
<tr>
<th>0. Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Preliminaries</td>
</tr>
<tr>
<td>2. Long Description</td>
</tr>
<tr>
<td>3. Abstract</td>
</tr>
<tr>
<td>4. Original Authors</td>
</tr>
<tr>
<td>5. Acknowledgements</td>
</tr>
<tr>
<td>6. Bib. Refs</td>
</tr>
<tr>
<td>7. Country</td>
</tr>
<tr>
<td>8. Classification</td>
</tr>
<tr>
<td>9. Variables</td>
</tr>
<tr>
<td>10. File Locations</td>
</tr>
<tr>
<td>11. Save and Curate</td>
</tr>
</tbody>
</table>

**Occupational Classification - Standards Used**

Please provide details of any custom classifications used in the editor under "Custom".

- [HISCO (5-digit)]
- [Add classification]

**Selected**  **Browse**  **Custom**
**Variables**

Please add a description of any variables that your data file(s) contain. For instance, the relevant information might be the variable name and value label items in something like an SPSS or Stata format data file. This section isn't compulsory, and it will not be applicable to all data resources deposited.

<table>
<thead>
<tr>
<th>Id</th>
<th>Short name (40 characters max)</th>
<th>Description</th>
<th>Remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HISCO</td>
<td>Historical Standard Classification of...</td>
<td><img src="remove.png" alt="Remove" /></td>
</tr>
<tr>
<td>2</td>
<td>HISCAM</td>
<td>HIS-CAM scale v1.1 (CAMSIS scale for the nineteenth century)</td>
<td><img src="remove.png" alt="Remove" /></td>
</tr>
</tbody>
</table>

<< first < previous

Save Cancel
DAMES Curation Tool
Curate a data resource into the DAMES Infrastructure

File Locations
Is the dataset to be curated stored within your personal DAMES folder or is it available from a remote website?

DANES Folder  Remote Web Site  Add File/Folder

<table>
<thead>
<tr>
<th>filename</th>
<th>Role</th>
<th>Remove</th>
</tr>
</thead>
<tbody>
<tr>
<td>ome/red5/My Data/demo/hiscam.ui.cda</td>
<td>Data File</td>
<td></td>
</tr>
<tr>
<td>ome/red5/My Data/demo/hiscam.ui.xls</td>
<td>Documentation</td>
<td></td>
</tr>
</tbody>
</table>

<< first  prev  next >>
Also automatically uploaded to searchable eXist database
Example: The Fusion Tool

- Scenario: A soc sci researcher wishes to fuse Scottish Household Survey data with privately collected study data:
  - Uses the data curation tool to upload the data
  - Uses the data fusion/imputation tool to select the data, identify corresponding variables, and to generate a derived dataset (held in the portal)
  - The metadata about this derived dataset is stored and (may be) made public through the portal
  - Another researcher can now search the portal (metadata) for SHS data and find the derived dataset

- DAMES metadata handling must facilitate this process
The Fusion Tool prototype

Select datasets (recipient and donor)

Select "common variables"

Select variables to be imputed

Select data fusion method

Submit to fusion "enactor"
Select datasets (recipient and donor)

Select "common variables"

Select variables to be imputed

Select data fusion method

Submit to fusion "enactor"
Select datasets (recipient and donor)

Select "common variables"

Select variables to be imputed

Select data fusion method

Submit to fusion "enactor"

DAMES Data Fusion Wizard

Stage 5: Output

Below is a list of the fusion settings used and a link to the file. Please confirm your settings and click the submit button to submit your job to DAMES. You can monitor the status of your job and download the resultant file from the job monitoring portlet. Thank you for using the DAMES Data Fusion Tool.

Job Submission Details

- Datasets
  - Donor: 1970 British Cohort Study
  - Recipient: Active People Survey

Metadata for result dataset
Fusion job flow description

- The fusion/imputation is potentially resource greedy
  - Generalise to using a pool of computing resources
- We use a Job Flow Description Language (JFDL) to submit the job to the computing resources pool
- The JFDL job description includes references to:
  - Input data sets
  - Processing steps and their relationships
  - Outputs
- This must be recorded in the DDI record for the resultant data set
  - We are automating the translation JFDL→DDI3
Summary/Work in progress

- The DAMES infrastructure is still very much in the development phase
- We have a prototype outline portal, with prototype curation and fusion tools
- We are firmly fixed on using DDI 3 as our metadata standard, but we are still:
  - Refining the JFDL
  - Refining the DDI3
  - Improving generation of DDI3 from JFDL
  - Improving searching and discovery of datasets

Thank you!