



## e-Framework Service Usage Model Name

- Name: **Grid-Enabled Micro-Econometric Data Analysis (GEMEDA) Service Usage Model**

## Version

- 2.5

## Version History

Version	Date	Author	Description	Organization / Project
2.5	Oct 2008	M. Argüello Casteleiro	Draft	NCeSS, University of Manchester
		Pascal Ekin		NCeSS, University of Manchester
		Simon Peters		School of Social Sciences, University of Manchester

## Rationale

The SUM presented here is about GEMEDA (Grid Enabled Micro-econometric Data Analysis) [1, 2], which is a Grid-based application for social sciences.

Although the use of Grid technologies is currently not common within social sciences research, many social scientists today want to investigate complex research questions that mean combining datasets from a variety of sources. However, there are difficulties in converting and combining the various data. GEMEDA makes use of Grid technology to integrate the data, computation and presentation elements of an empirical economic modelling process. Thus, GEMEDA has provided invaluable insights into many of the technical and methodological issues that need to be addressed when Grid-enabling social sciences datasets and developing Grid-based services. For example, GEMEDA allows addressing substantive issues, such as determining United Kingdom ethnic minority welfare by means of using a form of statistical data fusion developed in the poverty mapping literature.

## Classification<sup>1</sup>

<i>To be provided by the submitter:</i>				
<b>SUM Type</b>	<input checked="" type="checkbox"/> Domain	<input type="checkbox"/> CORE (a commonly recurring SUM; designation requires e-Framework Integrity Group approval)		
<b>Domain(s)</b>	<input type="checkbox"/> Learning & Teaching	<input checked="" type="checkbox"/> Research	<input type="checkbox"/> Administration	<input type="checkbox"/> Common Libraries
<b>Maturity</b>	<input type="checkbox"/> Immature	<input checked="" type="checkbox"/> Mature		
<b>Purpose(s)</b>	<input type="checkbox"/> Exemplar	<input checked="" type="checkbox"/> Application	<input type="checkbox"/> Modelling	<input type="checkbox"/> Toolkit
<b>XOR (exclusive "or")</b>	<input checked="" type="checkbox"/> Service Genres <input type="checkbox"/> Service Expressions			
<b>Development Status</b>	<input type="checkbox"/> Proposed	<input type="checkbox"/> Developmental	<input checked="" type="checkbox"/> Prototype	<input type="checkbox"/> Production
<b>Deployment Scale</b>	<input type="checkbox"/> Isolated <input type="checkbox"/> Ubiquitous			
<b>State Behaviour</b>	<input type="checkbox"/> Stateful <input type="checkbox"/> Stateless			

<sup>1</sup> See definitions of the Service Usage Model Classification Scheme categories and their allowable choices at: <http://www.e-framework.org/Services/ServiceClassificationScheme/ClassificationSchemeForSUMs/tabid/817/Default.aspx>

Transactional Behaviour	<input type="checkbox"/> Transactional and ACID	<input type="checkbox"/> Transactional but Non ACID	<input type="checkbox"/> Non-Transactional
Batch Behaviour(s)	<input type="checkbox"/> Individual	<input type="checkbox"/> Batch	
Time-Constraint Behaviour	<input type="checkbox"/> Hard Real Time	<input type="checkbox"/> Soft Real Time	<input type="checkbox"/> None
Service End Point	<input type="checkbox"/> Provider	<input type="checkbox"/> Requestor	<input type="checkbox"/> Transcoder (both requests and provides)
Authentication/ Authorization Dependency	<input type="checkbox"/> Auth-Dependent	<input type="checkbox"/> Auth-Independent	
Protocol Binding(s) (only applies to service expression-based SUMs)	<input type="checkbox"/> Web Service <input type="checkbox"/> SOAP	<input type="checkbox"/> REST <input type="checkbox"/> HTTP	<input type="checkbox"/> Other
<b>To be determined by the e-Framework:</b>			
<b>Status</b>	<input type="checkbox"/> Approved	<input type="checkbox"/> Placeholder <input type="checkbox"/> Unapproved	<input type="checkbox"/> Superseded <input type="checkbox"/> Withdrawn
<b>Confidence Level</b>	<input type="checkbox"/> High	<input type="checkbox"/> Medium	<input type="checkbox"/> Low

## Description

Empirical economic modelling using secondary information can be considered as consisting of three steps: *data handling*, *econometric computation* and *results presentation or visualization*. One might loosely term these the workflow of an applied economist, though steps one and three are often overlooked as a necessary part of the full process. GEMEDA (Grid Enabled Micro-econometric Data Analysis) [1, 2] is a Grid-based application for social sciences which demonstrates that it is now feasible to consider a computational environment that enables the full process by using the technology available from e-Science or e-Research (also known as *cyberinfrastructure* in the United States).

The application of Grid technology to the underlying substantive problem allows one to integrate all three elements of the empirical modelling process within a Web browser based interface suitable for use by a casual (i.e. non-expert) user. Ultimately a grid-enabled analysis will allow users to produce their own bespoke output, addressing a problem of interest to themselves.

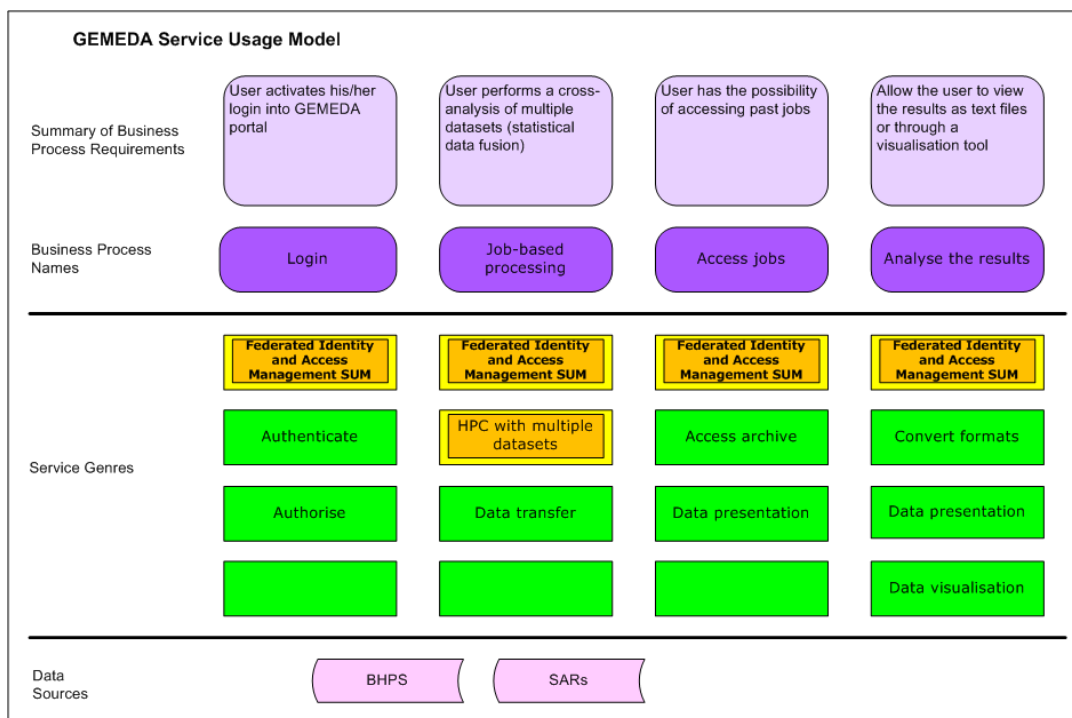
## Business Process Modelling

Based on [1, 2] and other relevant documentation and interviews material, the following business processes can be distinguished in GEMEDA:

- Single sign-on login into GEMEDA portal
- Authentication and Authorization
  - Certification
  - Allowance (= Permission)
    - Access to data
    - Access to computational resource
- Set-up and Launch a Job (allows to submit jobs to grid environments), where set-up a job implies to provide a “*Job description*” in terms of
  - Select the data parameters
    - *Select the economic variables for the analysis*

- *Select the geographic areas*
    - *Select the ethnic minority groups*
  - Select the algorithm parameters (econometric algorithm)
    - *Select the welfare measure to estimate*
  - Select the parameters specific to the computer resources (HPC)
    - *Select the HPC node and the number of processors to use*
- Job monitoring (monitor the progress of Grid operations – Show Job Status). Once a job has been launched, the user receive notifications about:
    - How the algorithm is being processed, i.e. about the code execution
    - The underlying infrastructure (i.e. the Grid) where the job is running
  - Access jobs (GEMEDA archives *past jobs* so they can be analysed or viewed a later stage), so it allows to access the “*Job description*” of a job done.
  - Analysing the results (Result Presentation or Visualisation)
    - View the results as text files
    - Launch a GIS visualisation tool (Java applet) which may imply a data transformation (convert formats) stage

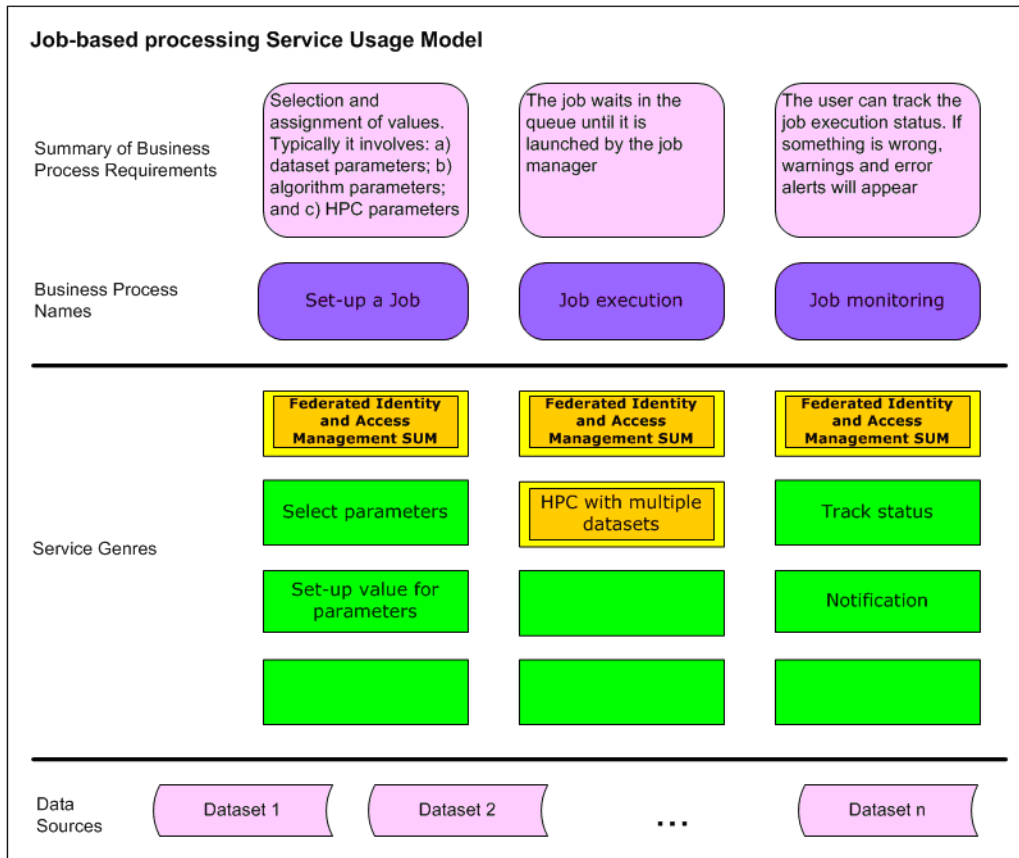
## SUM Diagram



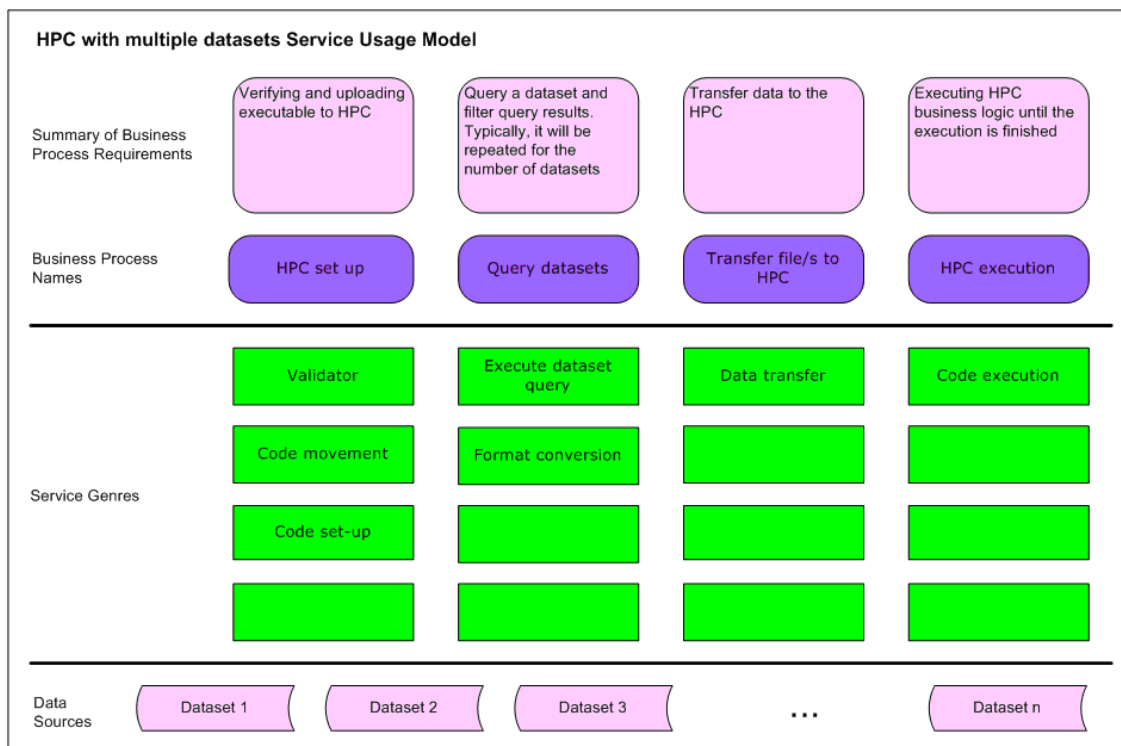
Visio® template for SUM diagram, revised 20070822  
 Template © Copyright 2007, e-Framework Partners

A modularise approach to SUMs was followed, and two embedded SUMs were identified:

- **Job-based processing SUM**
- **HPC with multiple datasets SUM**



Visio® template for SUM diagram, revised 20070822  
 Template © Copyright 2007, e-Framework Partners

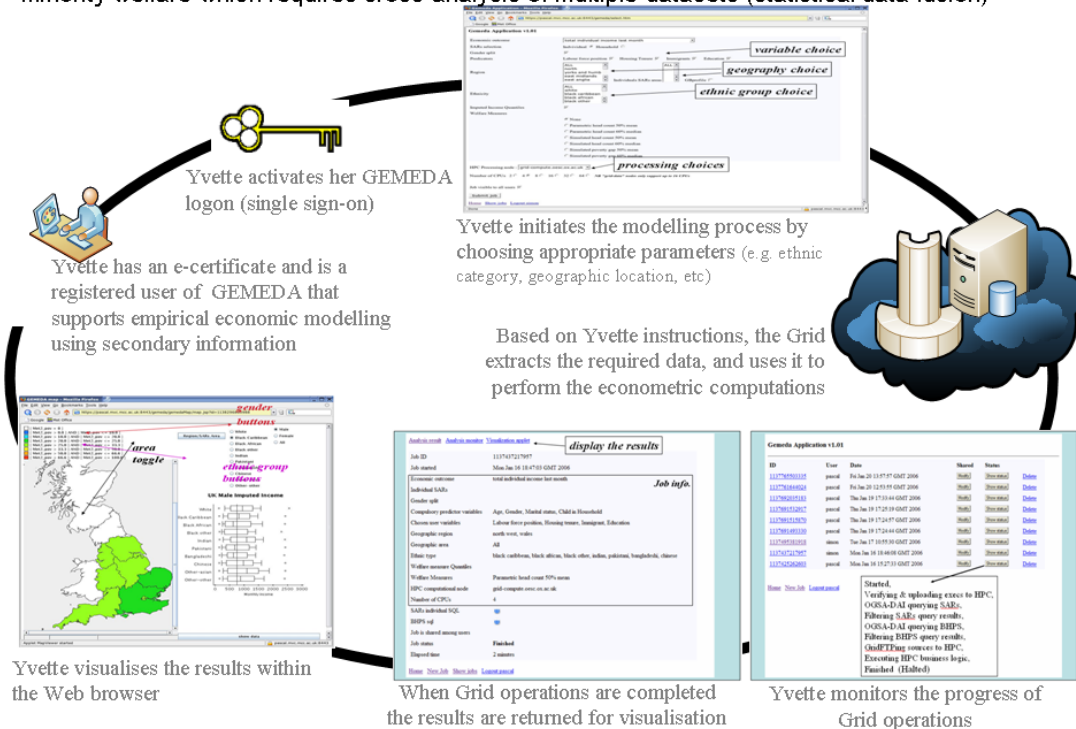


Visio® template for SUM diagram, revised 20070822  
 Template © Copyright 2007, e-Framework Partners

## Usage Scenarios [optional]

**GEMEDA (Grid Enabled Micro-econometric Data Analysis)** a Grid based demonstrator

Yvette is an econometrician that has an e-certificate. She wants to determine the UK ethnic minority welfare which requires cross-analysis of multiple datasets (statistical data fusion)

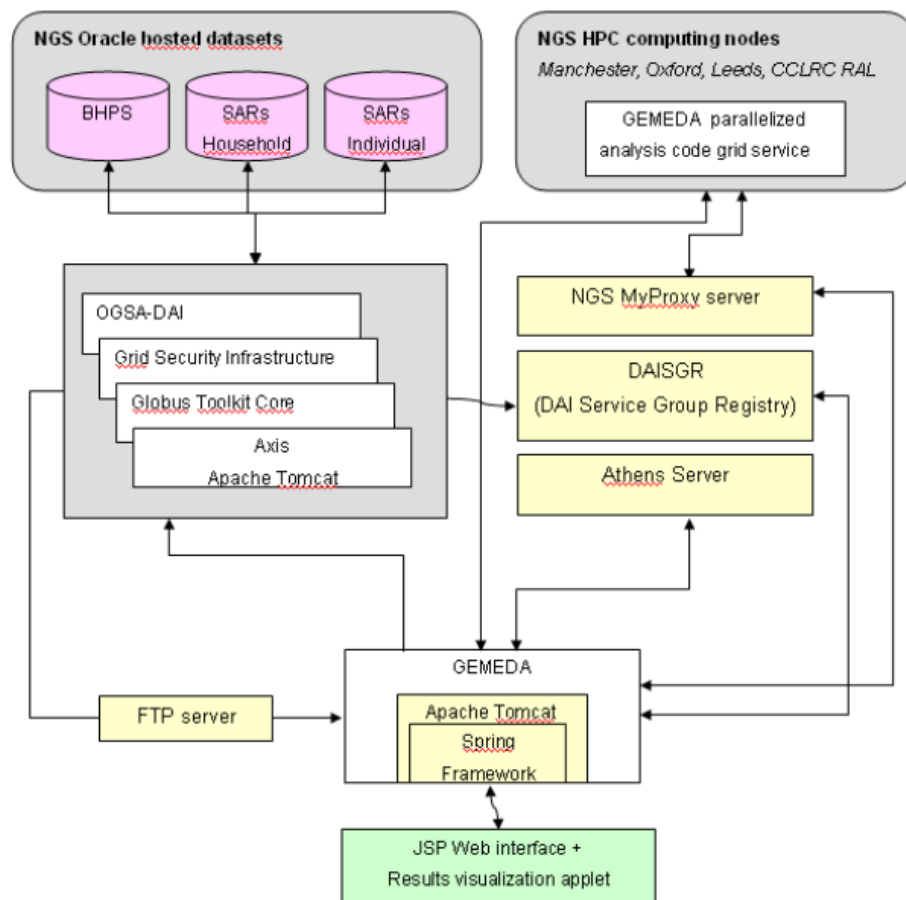


## Functionality

- **Single sign-on login into GEMEDA portal:**  
Users enter their user name & password. First time users, or users whose credentials need updating, will be redirected to the user's credentials page i.e. Athens credentials and the National Grid Service UK (NGS) MyProxy credentials (user chosen login name and pass-phrase).
- **Authentication and Authorization :**
  - Athens authentication and authorization is carried out for each session on the user's behalf.
  - GEMEDA employs MyProxy certificates throughout, by which means it can orchestrate a grid workflow.
- **Set-up and Launch a Job:**
  - Users enter job parameters through various menus, lists & bullet points.
  - Users select HPC machine (NGS core nodes) and the number of processors where the processing job will be carried out.
- **Job Execution -** Once a user has initiated a job run, the GEMEDA carries the following processes:

- GEMEDA generates the necessary SQL to query databases - data results are uploaded to the user's NGS account by means of the OGSA-DAI middle-ware using the GridFTP protocol.
  - GEMEDA verifies the compiled Fortran executable resides on the user's account and uploads the executable file if it does not exist.
  - GEMEDA generates a job specification file and uploads it to user's account space.
  - GEMEDA launches MPI job run through the Globus job broker.
- **Job Monitoring**  
The status of a job is accessible by the user through the use of Ajax technology (dynamic loading of content on the web page).
  - **Access jobs:**  
Job parameters and results are archived by the GEMEDA and can be made public for other GEMEDA users to consult.
  - **Analyzing the results (Result Presentation or Visualization):**  
Job results can be viewed and downloaded in text format or visualized, and further analyses, by means of an integrated GIS applet.

### Structure & Arrangement



**Figure 1.** Overview of the GEMEDA architecture

The user communicates securely (HTTPS) with the GEMEDA Web Service through the user interface – a web based light client. Various menus allow the user to choose between SARs individual and household data (hosted as an Oracle database) and the actions applicable for the selected data.

The GEMEDA service transforms the user choices in a series of SQL queries. The queries are sent to OGDADAI – a grid middleware component that gives authorised users access to disparate types of data storage mechanisms through a single common API.

The GEMEDA logic is a parallelized FORTRAN 95 executable launched by Globus as an HPC job. Once the job is completed, the results are written to file and downloaded through GridFTP by the web service. These results are then sent back to the user interface for viewing.

### ***Data Sources Used***

- The British Household Panel Survey (BHPS) provides the small scale survey data. BHPS is a longitudinal (panel) study with yearly waves.
- The Sample of Anonymised Records (SARs) provides the large scale Census data. SARs are a random sample of individuals and households from the UK Census. The project used 1991 data because of projected confidentiality restrictions on the publicly available version of the 2001 SARs.

### ***Services Used***

#### **List of Services Genres:**

- Authenticate [it appears in the list of e-framework Service Genres – 9 May 08]
- Authorise [it appears in the list of e-framework Service Genres – 9 May 08]
- Access archive
- Code movement
- Code set-up
- Code execution
- Convert Formats [it appears in the list of e-framework Service Genres – 9 May 08]
- Data Presentation
- Data transfer
- Data visualisation
- Execute dataset query
- Notification
- Select parameters / variables
- Set-up values for parameters / variables
- Track status
- Validator

## **CORE SUMs Used [recommended]**

The following Commonly Recurring (CORE) SUM has been reused from the e-Framework Service Usage Models Registry:

- Federated Identity and Access Management SUM [it appears in the list of e-framework Service Usage Models – 9 May 08]

## **References**

This document takes into account existing material that has been published as conference and journal papers. The most relevant papers are the following:

[1] Peters, S., Clark, K., Ekin, P., Le Blanc, A. & Pickles, Pickles (2007) '*Grid Enabling Empirical Economics: A Microdata Application*', in *Journal of Computational Economics*, Springer, vol. 30, no. 4, pp. 349-370

[2] Peters, S., Ekin, P., LeBlanc, A., Clark, K. & Pickles, S. (2006) '*Grid Enabled Data Fusion for Calculating Poverty Measures*', in *Proceedings of the UK e-Science All Hands Meeting*.

[3] M. Argüello, S. Peters, and P. Ekin: *Towards a collective knowledge base: sharing the expertise acquired on developing grid-based e-science and e-social science applications*. Online proceedings of Oxford e-Research Conference, Oxford, UK, September 2008

[4] M. Argüello, P. Ekin, A. Turner, S. Peters, P. Townend, M. Fraser, P. Halfpenny, R. Procter, A. Voss, and M. Jirotko: *Highlighting e-Infrastructure patterns in Grid-based e-Social Science applications*. Accepted for Regular Session at UK e-Science All Hands Meeting 2008 (AHM 2008), Edinburgh, UK, September 2008

[5] M. Argüello, P. Ekin, S. Peters, M. Fraser, P. Halfpenny, R. Procter, A. Voss, and M. Jirotko: *A case study about how e-Infrastructure is used within the Social Sciences*. Online proceedings of 4th International Conference on e-Social Science, Manchester, UK, June 2008



This SUM is licensed under:

Creative Commons Attribution-NonCommercial-ShareAlike 2.5 licence

<http://creativecommons.org/licenses/by-nc-sa/2.5/au/>

### **Submitting the Service Usage Model Description**

For additional guidance in preparing the Service Usage Model description, refer to [Guidelines for Submitting a Service Usage Model to the e-Framework](#)<sup>2</sup> and the technical definitions of the [Service Usage Model Description Elements](#).<sup>3</sup> For further assistance, contact the e-Framework editor at: editor@e-framework.org

Prior to submitting the description, please read the e-Framework [Intellectual Property Rights statement](#).<sup>4</sup> Also add your information to the copyright statements in the footer of this template.

By submitting this document, you agree to contribute this document under the [Creative Commons Attribution-NonCommercial-ShareAlike 2.5](#)<sup>5</sup> licence.

When you have completed the Service Usage Model description, go to the [Submit SUMs](#)<sup>6</sup> page at [www.e-framework.org](http://www.e-framework.org). Click on “Upload your submission” and follow the directions.

---

<sup>2</sup> Guidance for using this template: <http://www.e-framework.org/SUMs/SubmitSUMs/tabid/715/Default.aspx>

<sup>3</sup> Service Usage Model Element Definitions: <http://www.e-framework.org/Services/SUM/SUMElements/tabid/745/Default.aspx>

<sup>4</sup> Intellectual Property Rights Statement: <http://www.e-framework.org/About/Policies/tabid/611/Default.aspx>

<sup>5</sup> Creative Commons: <http://creativecommons.org/licenses/by-nc-sa/2.5/au/>

<sup>6</sup> Submit Service Usage Model template: <http://www.e-framework.org/SUMs/SubmitSUMs/tabid/715/Default.aspx>