(Abstract submitted to the UK e-Science All Hands Meeting 2011 on 2011-06-03)
A NeISS Collaboration to Develop and Use e-Infrastructure for Large-scale Social Simulation

Tom Doherty, Sam Skipsey, Andy Turner, John Watt

Abstract (of the abstract)
The National e-Infrastructure for Social Simulation (NeISS) project is attempting to develop e-Infrastructure to support social simulation research. This includes support for running contemporary dynamic demographic social simulation models developed in the GENESIS project. In these models, each individual human in a population is represented as a distinct digital entity and population change is driven by mortality and fertility rates that are stochastically applied. A portal based Graphical User Interface (GUI) to the model has been developed as a pair of standard JSR-168 portlets. One portlet is for specifying the model parameters and setting it running, the other is for reporting progress and allowing users to download results. A layer of programs enacted by the portlets stage data in and submit a job to a Grid computer that enacts the GENESIS model. Once the job is submitted some details are communicated back to a job monitoring portlet. Once the job is completed, results are stored and made available for download. Another portlet has been developed to manage the stored data. Collectively we call the system the GENESIS Simulator.

keywords
NeISS, e-Infrastructure, GENESIS, social, simulation, population, model, stochastic, portlet, grid

Social simulation is an attempt to model societies in a dynamical way in a digital computer [1]. “e-Infrastructure consists of social and technical arrangements around advanced, networked information and communications technologies that can enable new research practices and methods” [2]. This work is about the National e-Infrastructure for Social Simulation (NeISS) project, which is attempting to develop e-Infrastructure to support social simulation research [3].

Contemporary social simulation models have been developed as part of the Generative e-Social Science for Socio-Spatial Simulation (GENESIS) project. GENESIS is a second phase research node of the UK National Centre for e-Social Science funded by the UK Economic and Social Research Council (ESRC) for three years starting in October 2008 as a collaboration between the University of Leeds and University College London [4]. One focus of GENESIS work is to develop dynamic demographic social simulation models in which each individual person in a human population is represented as a distinct digital entity. The current population model is a-spatial and driven by mortality and fertility rates.

NeISS is a project that started in April 2009 and is funded for three years by the UK Joint Information Systems Committee (JISC) under its Information Environment programme [3]. NeISS has partners from 8 different institutions and was originally conceived to develop a general e-Infrastructure for social simulation. One part of NeISS work is to develop e-Infrastructure to support the GENESIS demographic daily time step social simulation models (henceforth referred to as the *model*). That work is the focus of this paper.

A portal based Graphical User Interface (GUI) to the model has been developed as a pair of standard JSR-168 portlets. One portlet is for specifying the model parameters and setting it running, the other is for reporting progress and allowing users to download results. The portlets are being developed and tested on on the DAMES Application Portal hosted at the National e-Science Centre (NeSC) in Glasgow

[5]. Being standard JSR-168 portlets means they can be readily migrated to other portal instances [6], but the model will only run with the other required programs in place. The next layer of programs enacted by the portlets are staging programs that organise the data and the core model executable and submit a job to a Grid computer. Once the job is submitted some details are communicated back to a job monitoring portlet. Once the job is completed, results are stored and made available for download. Another portlet has been developed to manage the stored data. Collectively we call the system the *GENESIS Simulator*.

The GENESIS Simulator is being used to produce results for the Leeds Local Authority District from 1991 to 2001. Annual resolution mortality and fertility rates are input to the simulation. These are calculated from Office for National Statistics (ONS) Vital Statistics on births and deaths and mid-year estimates of population as supplied to Paul Norman for ESRC Research Awards RES-163-25-0032 and RES-189-25-0162 [7,8]. These data are Crown copyright and are reproduced with permission of the Office of Public Sector Information. For each year, the model is run four times using different pseudo random number seeds to produce a range of results, one of which is selected to be used as input for the next year to be simulated. The selection is based on a comparison between simulated and input mortality and fertility rates. The simulation result with mortality and fertility rates closest to those input is the one selected. Some visualisation and discussion of these results are to be provided in the full paper.

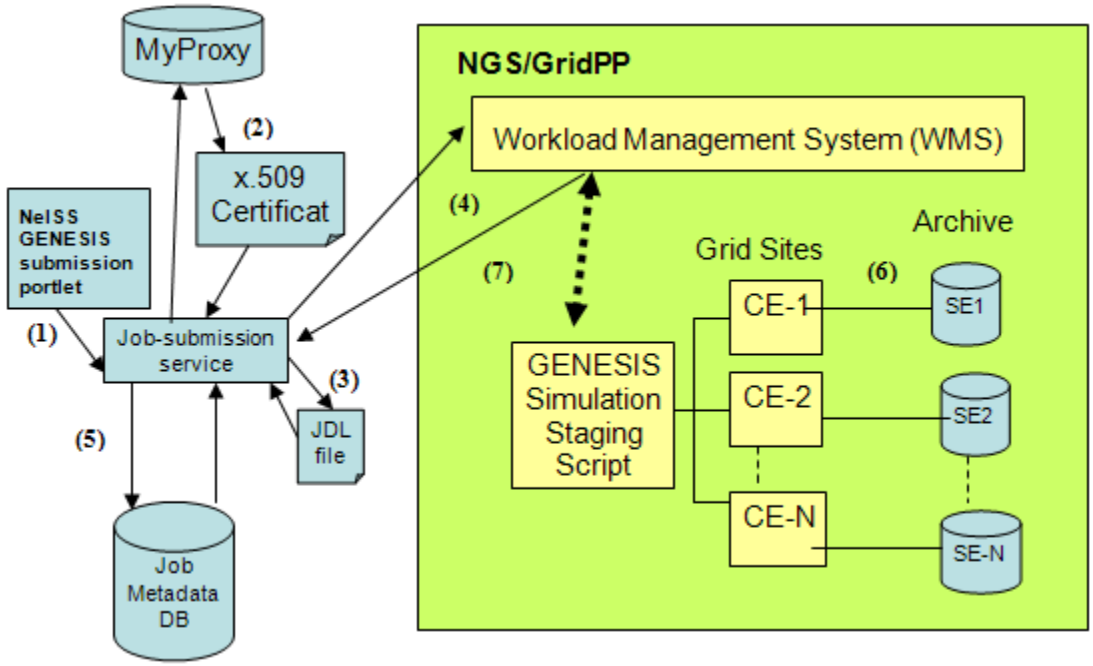
The first set of results are not expected to be used further although they will be compared with other available population data. Significant differences are expected, especially as the model does not currently explicitly model migration which is known to have a large effect on the population of Leeds [9]. The effect of net migration (the difference between the residential population moving out of and into a region) on other regions of the size of Leeds will vary, but in general, for larger regions with more population, the effect of net migration reduces as net migration becomes smaller relative to the size of the entire population. For this reason, results are also planned to be generated for England and compared. This will further test the e-Infrastructure and the demands on computational resources which are expected to be almost 100 times greater.

As the size of the population to be simulated increases and the number of steps in the simulation increases, the model demands more and more computational resource. A UK national simulation has requirements of input and output data in the size region of terabytes, and compute times for atomic model components that run for several days. The e-Infrastructure developed for the model will scale to this although the availability of large grid and cloud computing resources is not a certainty. Scaling up not only allows for more impressive results to be produced, but also allows for investigations into how he model scales. Computational demands are currently being met by resources at the National Grid Service (NGS) ScotGrid site and from GridPP [10,11,12]. The e-infrastructure uses the Virtual Organisation Membership Service (VOMS) solution [13] to handle access to NGS and GridPP resources. For this, a NeISS Virtual Organisation (VO) has been set up and has been approved by GridPP. The aim is to use the SARoNGS approach [14], but while this is being organised, effectively a single certificate is being used with an appropriate accounting system in place.

The e-infrastructure uses the NGS Workload Management System (WMS) to provide resource brokering-based job scheduling across all the Grid sites that support the NeISS VO [15,16]. The NGS user interface (UI) provides a gateway for WMS job submission to NGS and GridPP nodes via a command line interface [17]. VOMS proxy certificates allow users to submit Grid jobs using a Job Description Language (JDL) file that specifies the job parameters [18]. These parameters detail the type of file to execute on the worker node, abstract file paths for uploading data via the *input sand box*, and abstract file paths for retrieving data from the *output sand box*. Currently, the job staging script is executed on a single worker node and this script handles the running of the model. This script also pushes the results into a data archive

organised as Logical File Catalog (LFC) [19]. Some result files are also made available to the user via the output sandbox. Normally a user would submit their job and associated JDL file to the WMS using a *glite-wms-job-submit* command and also associate a delegated VOMS proxy. On successful submission, a job id is returned which is used to monitor the job status using the *glite-wms-job-status* command. Ordinarily, a user would fetch the job output when completed using the *glite-wms-job-output* command. All this command line work is now hidden from the user behind the simple and easy to use portal based GUI. Figure 1 depicts the current GENESIS Simulator workflow.

Figure 1. The basic GENESIS Simulator workflow



(1) After logging in and authenticating using shibboleth [20] the user initiates job submission via the portlet and the portlet then invokes the job-submission service
(2) Job submission service pulls user's proxy from MyProxy service [21] and creates VOMS proxy
(3) Job-submission service creates JDL file drawing in user input provided via the portlet interface
(4) jLite [22] API used for Java representation of glite-WMS commands – job submitted to WMS
(5) Job ID and associated Job metadata stored in Job metadata database using JDBC API [23]
(6) Model output saved in archive using WLCG tools and registered in LFC
(7) Model result metadata saved as output from worker node and passed back to portlet. Including GUID associated with population file (saved in archive) so that this file can be used as input to future jobs.

The complexity of Grid middleware coupled with grid certificates has proven to be a barrier to entry for researchers in some disciplines [24]. As portals were already being used within the NeISS project, it was fitting to develop and use portlets to facilitate job-submission, status-monitoring, output retrieval and job output comparison. The OMII SPAM-GP Shibboleth module [25] is used to login to the portal framework, and an iFrame [26] within one of the portal pages provides a link to the NGS Credential Translation Service [27] GUI where the user may generate their SARoNGS certificate, which can then be downloaded. A Registration Portlet allows the information provided in the portal account and the generated SARoNGSs proxy to be used to contact the VO administrator to request membership. Job metadata and monitoring is recorded in a database which makes it possible to track and manage each job submitted. A Management

Portlet allows for the deletion of results data stored in the database and the associated archived data. The main Job Submission Portlet is designed with a 'wizard' type flow where the user enters the necessary information for the simulation in a step by step fashion. The certificate and JDL configuration is transparent to the user.

Archive data management leverages the existing WLCG [28] gLite [29] infrastructure, in order to reduce the amount of additional work needed to support it across potential sites . Files generated by a job are stored at the local Storage Element (for the UK, most often a DPM [30] in front of a disk pool), and registered in the UK LFC at RAL [31]. Later jobs can be directed to sites holding local copies of required data, and copies of data can be replicated at other sites, as the files are managed entirely in terms of their Globally Unique Identifier (GUID) [32] assigned by the LFC.

Via the portlet interfaces, the model is relatively easy to use. In the full paper and presentation we detail the implementation outlined above, present some initial results, and consider a road map for sustaining our collaboration.

References and further information

1. Gilbert N., Troitzsch K.G. (2005) Simulation for the social scientist. Second edition. Milton Keynes: Open University Press. ISBN-13 978 0335 21600 0.
2. Voss A., Vander Meer E., Fergusson D. (2009) Research in a Connected World. http://www.researchconnect.org/book
3. GENESIS http://www.genesis.ucl.ac.uk/
4. NeISS http://www.neiss.org.uk
5. DAMES Applications Portal https://dames.nesc.gla.ac.uk/
6. Java Specification Request 168: Portlet Specification http://www.jcp.org/ja/jsr/detail?id=168
7. What happens when international migrants settle? Ethnic group population trends and projections for UK local areas under alternative scenarios http://www.esrc.ac.uk/my-esrc/grants/RES-163-25-0032/read
8. Ethnic group population trends and projections for UK local areas: dissemination of innovative data inputs, model outputs, documentation and skills http://www.esrc.ac.uk/my-esrc/grants/RES-189-25-0162/read
9. Wu B.M., Birkin M.H., Rees P.H. (2008) A spatial microsimulation model with student agents. Journal of Computers, Environment and Urban Systems, volume 32, pages 440-453. DOI: 10.1016/j.compenvurbsys.
10. NGS http://www.ngs.ac.uk
11. GridPP http://www.gridpp.ac.uk/
12. ScotGrid http://www.scotgrid.ac.uk/
13. R. Alfieri, R. Cecchini , V. Ciaschini , L. dell'Agnello, A. Frohner, A. Gianoli , K. Lorentey , and F. Spataro (2003) VOMS, an Authorization System for Virtual Organizations. https://twiki.cnaf.infn.it/twiki/bin/viewfile/VOMS/WebDocumentation?rev=1;filename=VOMS-Santiago.pdf
14. SARoNGS http://www.jisc.ac.uk/whatwedo/programmes/einfrastructure/sarongs.aspx
15. Workload Management System http://glite.web.cern.ch/glite/wms/
16. VOMS admin for VO: neiss.ac.uk https://voms.ngs.ac.uk/voms/neiss.org.uk
17. NGS WMS UI http://www.ngs.ac.uk/ui-wms
18. JDL Specification https://edms.cern.ch/document/590869/1
19. Logical File Catalog http://www.gridpp.ac.uk/wiki/LCG_File_Catalog
20. The Internet2 Shibboleth framework http://shibboleth.internet2.edu
21. MyProxy http://grid.ncsa.illinois.edu/myproxy/
22. jLite Java gLite API http://code.google.com/p/jlite/
23. Java Database Connectivity (JDBC) API http://www.oracle.com/technetwork/java/javase/tech/index-jsp-136101.html
24. Jensen J., Spence D., Viljoen M. (2007) A Scalable PKI for a National Grid Service. http://middleware.internet2.edu/pki07/proceedings/11-jensen-pki-national-grid.pdf
25. SPAM-GP http://www.nesc.gla.ac.uk/projects/omii-sp/index.html
26. iFrame http://en.wikipedia.org/wiki/HTML_element#Frames
27. NGS CA Hierarchy http://wiki.ngs.ac.uk/index.php?title=NGS_CA_Hierarchy
28. Wordlwide LHC Computing Grid (WLCG) http://lcg.web.cern.ch/LCG/
29. gLite Lightweight Middleware for Grid Computing http://glite.cern.ch/
30. Disk Pool Manager (DPM) http://www.gridpp.ac.uk/wiki/Disk_Pool_Manager
31. The Rutherford Appleton RAL http://en.wikipedia.org/wiki/Rutherford_Appleton_Laboratory

32. Globally Unique Identifier (GUID) http://en.wikipedia.org/wiki/Globally_unique_identifier