IntegraEPI: a Grid-based Epidemic Surveillance System

Fabricio A. B. da SILVA[a,1], Henrique F. GAGLIARDI[a,b], Eduardo GALLO[a], Maria A. MADOPE[a], Virgílio C. NETO[c] , Ivan T. PISA[c], Domingos ALVES[d]
[a]*Programa de Mestrado em Informática,UNISANTOS, Santos, Brazil*
[b]*LCCASC,UNISANTOS, Santos, Brazil*
[c] *Departamento de Informática em Saúde (DIS), UNIFESP, Brazil,*
[d]*Universidade de São Paulo, Ribeirão Preto, Brazil*

Abstract: The proposal of new analytical techniques has guided innovative methodological developments in public health interventions. The goal of this work is show advances in the development of a large scale system for space-time visualization, monitoring, modeling and analysis of epidemic data using a Grid platform. The resulting virtual laboratory, dubbed IntegraEPI, is expected to provide better epidemic forecasting and to define better strategies to fight the spread of a disease, in which new population-level interventions could be evaluated and iteratively refined using computational simulations, with tangible benefits for real-world epidemic prevention and control efforts.

Keywords: Grid Computing, Epidemiology, Simulation, Data Integration

Introduction

Conventional epidemiology of infectious disease requires extensive collections of population, health and disease patterns data, as well as data related to environmental factors and social conditions. An epidemiologic study may focus on a particular region or a particular outbreak, or it may take as its theme the epidemiology of a condition across a wide area. The range and amount of data required will, therefore, vary depending on the type of study. Moreover, lack of data quality control, lack of definition about the contents to be stored, storage heterogeneity and resource availability are some problems that must be solved to allow more precise and thorough studies in epidemiologic vigilance. Furthermore, analytical studies to identify risk factors related to epidemic development are eventually used by health agencies [1][2]. These studies need several types of data, such as geo-referenced disease cases, space-temporal environmental data relevant to the epidemic prevention and population data based on demographic and geographic information with territory expressiveness.

Considering this scenario, we present in this paper some advances in the development of a large scale system for space-time visualization, monitoring, modeling and analysis of epidemic data studies using a Grid platform [3]. This system, dubbed IntegraEPI, is capable to provide the integration of heterogeneous databases related to epidemic analysis and to make available analytical and computational methods to

[1] Correspondence to: Fabricio Silva, Rua Dr. Carvalho de Mendonça, 144, Santos, SP, Brazil (e-mail: fabricio@unisantos.br). The authors acknowledge CNPq, CAPES and FAPESP for their financial support.

increase the predicting capability of the public heath system, in order to optimize its activities and resources when dealing with epidemic outbreak and prevention.

Particularly, in this work we concentrate ourselves in the data available in two areas in Brazil: the São Paulo State metropolitan coastal region known as "Baixada Santista", composed of nine cities, and a northeastern city of São Paulo State named Ribeirão Preto. Thus, all the conceived systems and infrastructures implemented were initially tested for these Metropolitan regions, carefully incorporating the diversity of the geographic spatial characteristics of the micro-regions. In order to be able to show the system capability in monitoring an epidemic, we work with the local public health system notification data. In this paper we focus on epidemic models for the Dengue fever which is a common disease in Brazil.

This paper is organized as follows: the need for a grid infrastructure is discussed in section 1. The general IntegraEPI architecture is presented in section 2. The services that compose the architecture are detailed in subsections 2.1, 2.2 and 2.3, and section 3 contains our final remarks.

1. Why developing a Grid-Enabled System?

On the beginning of this project many technological decisions were made and the choice of developing grid-based applications was made considering the real nature of the class of applications we should deploy.

In fact, the first feature which IntegraEPI would have to implement is the capability to integrate several health databases containing epidemiologic information with statistical and geographic databases. Such databases contain the necessary data to detect the patterns present on the disease notification process, considering external factors like socio-economic and environmental conditions. The use of grid services to integrate these databases provides transparent and simplified access for geographically distributed databases as if they would be a single and unique virtual database.

Another major reason for developing the IntegraEPI modules as grid-based applications is the forecasting capability provided by the epidemical model simulations, which would require large computational power. The InteraEPI simulator module is a parameter sweep application, and this type of application is especially well-suited for the Grid. The large number of simulations and the amount of computing power needed justify the Grid as the platform of choice to implement the system.

In front of this scenario we have chosen the Globus Toolkit 4 (GT4) middleware [12] to deploy a Grid platform over which the IntegraEPI modules would be implemented. The Globus Toolkit 4 is composed of several services designed for computational grids which provides, for instance, transparent data access [6] and resource virtualization [3][4] in a geographically distributed heterogeneous system. In the next section we should discuss the IntegraEPI system architecture and its main features.

2. The IntegraEPI Architecture

The IntegraEPI system architecture is based on the Open Grid Services Architecture (OGSA) specification [5] and is organized as shown in Figure 1. In fact, there are three main services which were developed: the data integration service Integra-GISE

(IntegraEPI-Grid Data Integration Service) [11] [10], the simulation service Integra-Model and the analysis service Integra-Analysis [9].

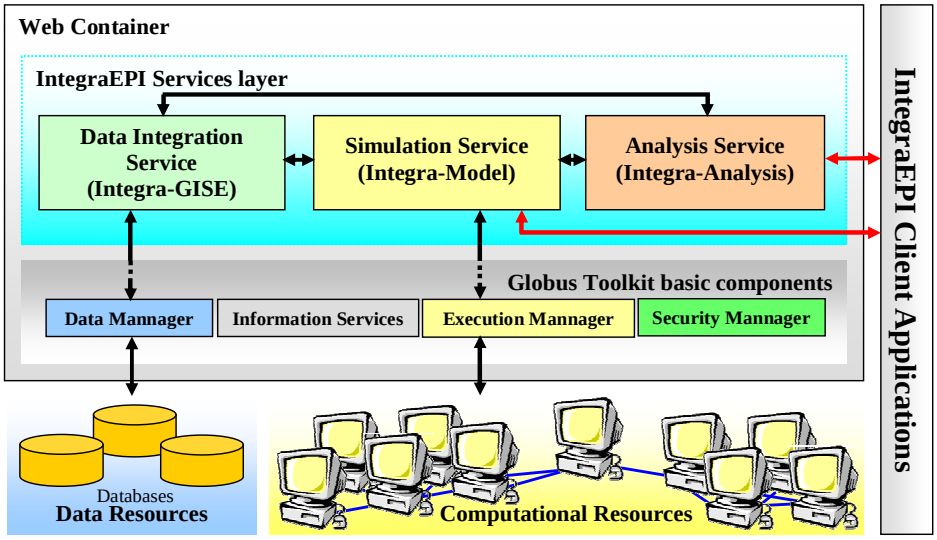

Figure 1. The Integra-EPI System Architecture.

Each one of the IntegraEPI services cover an important aspect of an epidemic surveillance system, such as: the integration of heterogeneous, geographically distributed populational and disease notification databases, a simulation service based on epidemiological models for prediction of the forthcoming trends related the transmission of a particular disease over a city population and, at last, statistical tools to analyze and visualize the status of such diseases, which make possible the inference of risk indicators and to establish epidemiological thresholds to trigger alerts if some risk situation is detected.

Due to the interoperability provided by the use of grid services and the definition of a Common Data Model, the IntegraEPI services are able to interact with each other. The integrated data provided by the IntegraEPI-GISE may be used by both the simulation service Integra-Model and the analysis service Integra-Analysis for estimator inference and detection of risk situations. At the same time, the simulation module can use any of the analysis tools provided by Integra-Analysis to study simulated patterns. Such interoperability allows a large degree of flexibility for service composition and implementation.

In the same figure we can see (on the middle) the Globus Toolkit components which provide services used by IntegraEPI modules such as reliable file transfer, replica location service, security and authorization managers, execution manager components, information services and common runtime components.

After this brief system description, we should describe, in the following subsections, the IntegraEPI main components in more detail.

2.1. Grid Data Integration Service (Integra-GISE)

The data integration service INTEGRA-GISE was developed to obtain data from multiple data sources through a single point of access. Particularly, this service provides an efficient management of the available computational resources. It also makes use of the grid security infra-structure and unique login to access multiple data sources.

Similarly as presented in the general IntegraEPI architecture, the Integra-GISE architecture is divided into layers composed by grid services and resources. The Figure 2 illustrates the GISE architecture, its layers and components.

In the simulation and analysis modules, the IntegraEPI system needs to access different data types like maps, epidemic data, geographic information and social-economics data. The required data are geographically distributed in many DBMS such as PostgreSQL, SQL Server, and Oracle. The GISE service provides the necessary data to other IntegraEPI modules in a transparent manner, by implementing a set of grid data services. In particular, several problems were considered when designing and implementing this module:

a) The number of sources of data is large, complicating still more issues related with the conflict resolution.

b) The available data sources vary dynamically. Therefore the addition and removal of data sources should be made with minimum impact in relation to integrated schemas.

c) The sources of data can have different computational capabilities. The sources of data to be integrated through a grid can vary of simple archives up to parallel DBMS.

For the Integra-GISE we have used the OGSA-DAI (Open Grid Service Architecture - Data Access Integration) [8] services (provided with GT4). The OGSA-DAI services provide wrappers to access data from different sources.

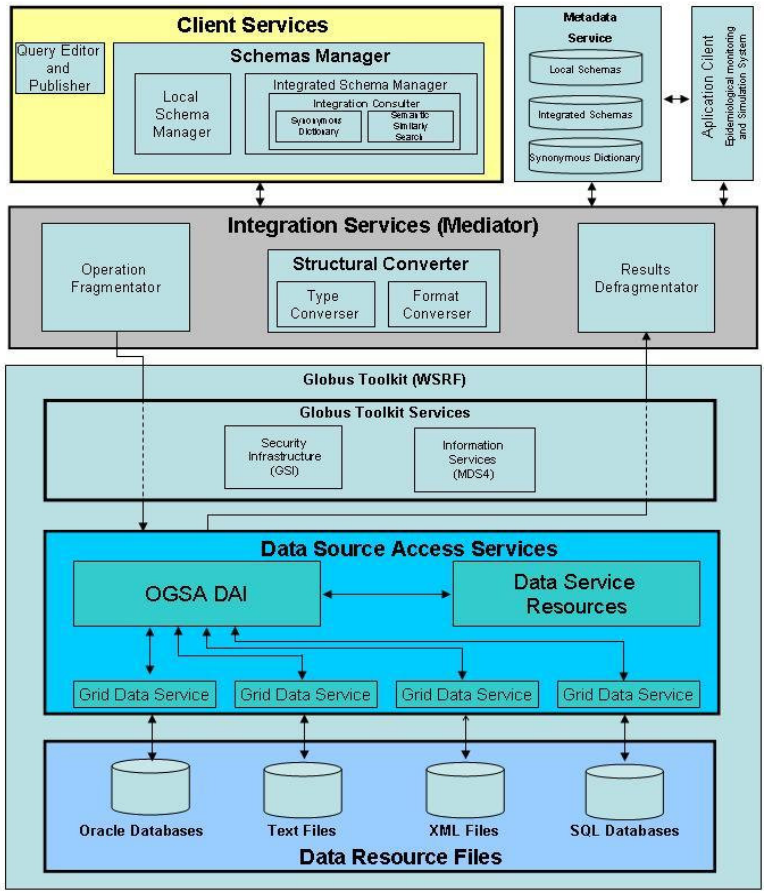

Figure 2. The layers and components of Integra-GISE architecture.

Together with the Integra-GISE service we have also implemented a Metadata Catalogue Service (MCS) (which also access underlying databases using OGSA-DAI services) to provide information about the system data sources which are available through database schemas. Two types of schemas compose the Canonic Data Model of the Integra-GISE service: local data base schemas (GISE-LOCAL-SCHEMA) and integrated schemas (GISE-INTEGRATED-SCHEMA). The integrated schemas may reference several local data base schemas.

The Integra-GISE client layer supplies the necessary components and interfaces used by final users. The layer is composed of the local schema manager, integrated schema generator and the query editor. Furthermore, in this layer it is also implemented the Integration Helper, which is used by the user to solve syntactic conflicts in the process of generating integrated schemas. It is worth noting that, in the Integra-GISE architecture, integrated schemas are generated manually.

The Client Layer builds the query operation through the Query Editor and Publisher, forwarding it to the Integration Services Layer (Mediator). The mediator represents the key element of this integration architecture since it is responsible to split

a single query into the many subqueries needed to access the diverse local databases which were referenced by the GISE-INTEGRATED-SCHEMA. The mediator is also responsible for gathering the partial results into a single result set.

When the client application submits a query to the Data Integration Module, some operations are required to process the query. The first step for query resolution is to fetch into the MCS all the information (schemas) related to the local databases referenced at the GISE-INTEGRATED-SCHEMA.

Therefore, the Integra-GISE Service first invokes the MCS service for obtaining the Integrated Schema. The Integrated Schema is a XML-based document containing information about the related databases such as a brief description of them, the databases ID, the databases alias, as well as in which fields the integration occurs.

After the selection of the local databases associated with the query, some specific information is still needed to build all sub-queries with the correct syntax. The Data Integration Service then fetches the description of each individual database (GISE-LOCAL-SCHEMA) in the MCS. Each GISE-LOCAL-SCHEMA is a XML document describing an individual data source.

All the subqueries are executed through the proper data access services and the resulting data sets are joined on a defragmentation process, which can also use the Synonym Dictionary to resolve any database structural or data type conflict, and so the resultant integrated data set is delivered as a single XML Document to the caller application.

Therefore, environmental, socio-economical, epidemiological, biological, weather and relief data would be retrieved by this service which virtualizes the access to several databases on the grid. Thus, the Integra-GISE importance relies on the data gathering mechanism needed to provide the necessary data for Integra-Analysis and Integra-Model services, which will be presented in the following subsections.

2.2. The Analysis Service (Integra-Analysis)

The analysis module Integra-Analysis is used for spatio-temporal data analysis, through a friendly user interface which provides considerable analysis flexibility. This analytical tool, in the first place, was developed to support a whole set of visualization tools, methods of spatio-temporal cluster detection per area of data, analysis methods of temporal series (self-correlation and spectral analysis) and methods of spatial interpolation for environmental and by area data. Figure 3 shows a particular analysis aimed at the detection of “hot zones” of Dengue spreading in Ribeirão Preto city. The distribution of urban Dengue fever cases over the city map with several clusters of infected individuals can be visualized in (a). To improve the analysis process we can use the *quadrat analysis* method in (b), in which the city map is divided into squares to visualize the ‘hot regions’ of the city where the epidemic spreading is more critical. After identifying those regions, the analysis module is also capable in (c) to display the most suspicious risk places for mosquito reproduction in the neighborhood (Military Police headquarters, junkyards, airports and cemeteries, for instance). Another visualization tool can be used (d) to identify directly over a real city image the most affected area of that cluster and the possible places of disease spreading due to historical data or behavioral analysis, which helps health agents to act in a more accurate way in the Dengue counter-attack.

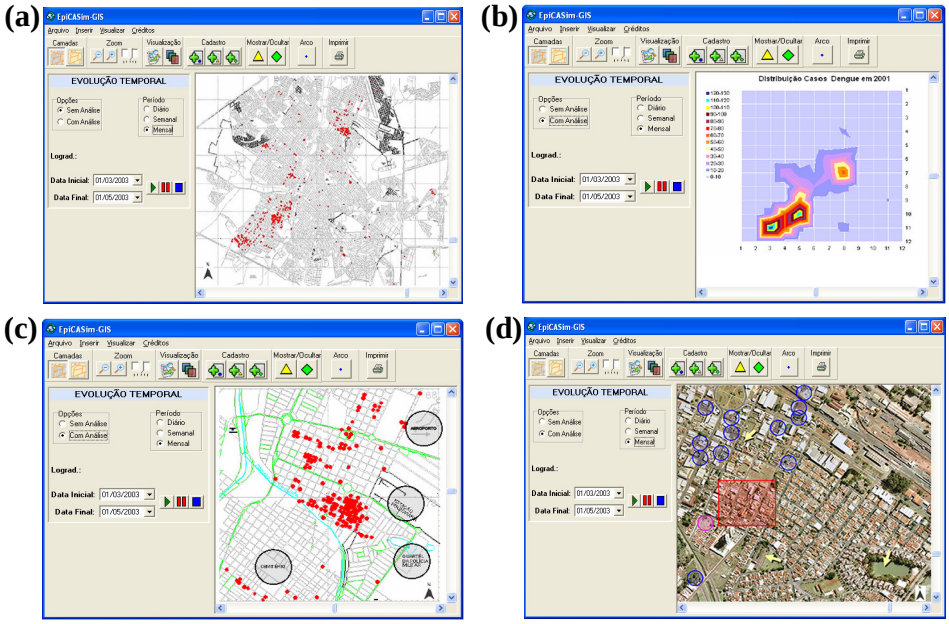

Figure 3. The analysis module features visualized in a client application. This figure illustrates (in a clockwise direction) a monitoring analysis of Dengue cases in Ribeirão Preto, a "hot zone" analysis, an analysis considering strategic points (junkyards, unused terrains) and special landed properties (museums, clubs, public places), and the visualization showing real Dengue fever cases.

In a second moment these analysis tools may used interactively for calibration of both the simulation model and to define thresholds related to alarm and risk indicators, Dengue epidemic dissemination and population social vulnerability. Moreover, this module is able to construct environmental indicators with a scoring methodology to stratify areas in the cities by different levels of risk for Dengue occurrence and transmission.

2.3. The Simulation Service (Integra-Model)

The main goal of the simulation module is to act a forecasting tool for disease spreading through the simulation of epidemiological models based on individual-based networks [7]. This module is used to test the efficacy of several control measures combining richly structuralized GIS networks of the municipal districts, describing the urban structure in multiple scales. It also considers realistic estimates and parameterized populational mobility and interactivity models, as well as disease progress among its hosts.

Generally, this service is capable to collect information through the Integra-GISE integration service and build a virtual city where a disease model will be applied to verify the possible scenarios of disease spreading among a virtual population. In this way, the provided data is processed to define the address location of every infected person in the population network. The additional infrastructural information gathered is used to infer the susceptibility of the population by populational zones, considering poverty, educational level, socio-economical data as well to build the simulation

scenario. Therefore, the most important feature presented by this simulation service is the capability to provide a substrate to represent the reality of a given municipal district as a "virtual city".

In this world build from the information provided by Integra-GISE, aspects like weather, relief, demographical density, social caress indexes, historical epidemiologic data, hydrography, urban or rural limits and places with large probability or tendency for the development of some type of vector agent are considered. The data are transformed into model parameters to be submitted as a Bag-of-Tasks[2] (BoT) job on the grid, via the Globus Toolkit service WS-GRAM.

It is worth noting that the Integra-Model service is capable to adapt itself to the reality of a given city, just needing to be parameterized with the data (fetched by Integra-GISE) of the metropolitan region under study, since the inter-individual interactions modeled remain unaltered.

Thus, with a single model we can simulate the behavior of a disease considering various distinct scenarios, and every single disease will have its simulator which needs to be built in a modular and parameterized way. The fact is that every infectious disease has its particular rules, transitions and modus operandi. Therefore, a specific canonic data model is necessary for every disease model provided on Integra-EPI system.

The simulator itself is a parameter-sweep application, in which an independent task is generated for each different set of parameters, in a *bag-of-task* approach. The model for the dengue fever is in advanced stage of development and it is being used for testing the Integra-Model service. The dengue fever model is described in the following.

The main assumption of our approach is that the dengue transition rules are defined by two sets of states, each one representing the behavior of a distinct population. In the language of state-variable models, the humans are defined by a SEIR (Susceptible, Exposed, Infective and Recovered) model which, when in the exposed state, the individual is infected but still not infective [7]. The mosquito population is represented by a different SEI (Susceptible, Exposed and Infective) model as shown in Figure 4.

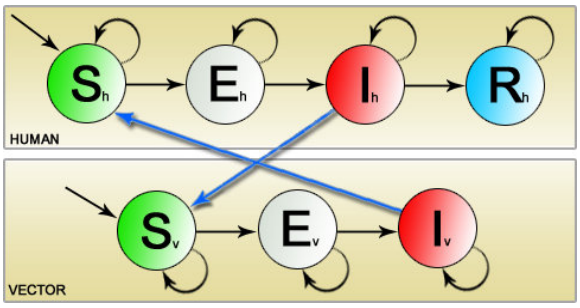

Figure 4. The schematic model of dengue spreading representing the stages of the disease for both populations, where *thick edges* represent the interaction among populations (mosquito bite) and the thin ones the internal state transitions in each population.

[2] In this work we use the terms "Bag-of-Tasks" and "Parameter-Sweep" interchangeably.

The dengue model has a singular characteristic, different from other models based in cellular automata: the use of two overlapping interacting automata cellular grids to represent the human and the vector (Aedes aegypti mosquito) populations. Therefore, the neighborhood of each individual has a different meaning and is not defined around each individual cell in the same population but at an equivalent position at the overlapping population, i.e. a neighborhood of a human cell is defined by the mosquitoes (vector) population and vice-versa. This special characteristic allows for the possibility of a human host both being locally infected by a vector and to infect another mosquito. This interaction (a bite), however, never occurs directly among the neighbors of a same population, because a human may only become infected by the Aedes aegypti bite and a mosquito only becomes infected by biting an infective human. Therefore, we propose a novel framework to model the spread of a dengue outbreak. Besides defining internal state transition rules for each iterative population (human and mosquitoes), we define iterative rules between these two cellular automata (Figure 5), reflecting the interaction between populations.

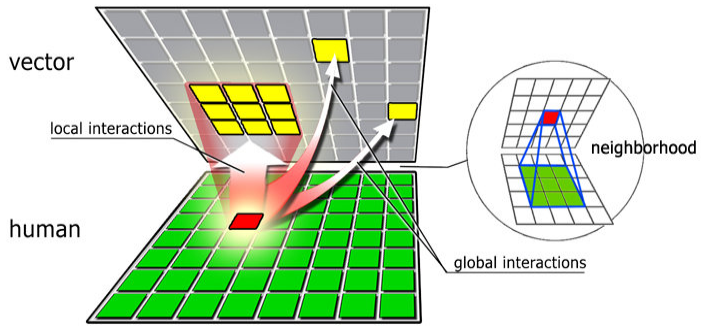

Figure 5. The local and global effects are shown in this figure. The *pointed squares* represent the mosquitoes affected by the local and global human infective influence. The same type of effects occurring in this bottom-up direction for human-vector interactions also occurs for the vector-humans interaction in a top-down direction at each simulation time-step. We can see also in this figure a schematic representation of the *neighborhood* of a single element.

To build the virtual city used in simulations several interrelated layers are used to map city features, like the demographic density, relief, hydrography and weather information. Therefore, considering the urban Dengue fever model [7], each population (human and mosquitoes) should have their own characteristics modeled. However, it is worth noting that some layers are more useful to humans than to mosquitoes and vice-versa. In general, using the Integra-Model service, we expect to be able to identify different levels of risk for a particular disease occurrence and transmission, for a predefined city or metropolitan area, considering the populational groups living in the city. Our model serves as a local strategic complement to other simulation models developed to identify epidemiological interactions within a given county or city.

3. Concluding Remarks

The essential proposal of this research project is to contribute for the modernization of the epidemiologic monitoring system by comparing results of detailed simulations with the observed experimental data related to the spreading of a disease, considering both temporal and geographic aspects. Particularly, the implementation over a computational grid platform open completely new perspectives for gathering data on large populations and - as a consequence - allow stratification of large scale Metropolitan epidemiology studies. Other advantages of this technology are the small deployment cost and high processing and storage capacities. These advantages become even more important when considering the deployment costs of mainframes or supercomputers for countries like Brazil.

Finally, it is important to emphasize that the expected results to be obtained during the development of this project do not apply solely to epidemics. There is a whole class of the public health problems of spatial and temporal nature, over which simulating, detecting, monitoring and visualizing patterns is part of the response to the problem.

4. References

[1] A. Wagstaff. *Socioeconomic inequalities in child mortality: comparisons across nine developing countries*. Bull World Health Organ 2000; 78:19-29.

[2] O. Axelson. *The character and the public health implications of ecological analyses*. In: Disease Mapping and Risk Assessment for Public Health (A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel & R. Bertollini, ed.), pp. 301-309, New York: Wiley and Sons, 1999.

[3] I. Foster. *What is the Grid? A Three Point Checklist*. GRID today, vol. 1, no. 6, 2002. Available at: http://www.gridtoday.com/02/0722/100136.html

[4] I. Foster, C. Kesselman, S. Tuecke. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International J. Supercomputer Applications, 15(3), 2001.

[5] I. Foster, C. Kesselman, J. Nick, S. Tuecke: *The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration*. June/2002. Available at: http://www.globus.org/research/papers.html

[6] I. Chervenak et al. *The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets*. Journal of Network and Computer Applications, 23:187-200, 2001.

[7] H. F. Gagliardi, F. A. B. da Silva and D. Alves. *Automata Network Simulator Applied to the Epidemiology of Urban Dengue Fever*. Springer-Verlag Berlin Heidelberg: Lecture Notes in Computer Science - ICCS 2006, Part III, LNCS 3993, pp. 297 – 304, 2006.

[8] *Open Grid Services Architecture Data Access and Integration*. Available at: http://www.ogsadai.org.uk/.

[9] A. Rezende, H. F. Gagliardi, R. C. Serafim, F.A.B. Silva, D. Alves. *IntegraEPI-GIS: A Geographic Information System to visualize and analyze the spatio-temporal patterns of the spread and control of epidemics.* Proceedings of the Spatial Epidemiology Conference 2006, London

[10] E. Gallo, H. F. Gagliardi, V. C. Neto, D. Alves, F.A.B. Silva. *GISE: A Data Access and Integration Service for a Grid-Based Epidemic Surveillance System*. Proceedings of the 40th Annual Simulation Symposium, 2007

[11] V. C. Neto, H. F. Gagliardi, A. Rezende, E. S. Gonçalves, E. Gallo, F.A.B Silva, I. T. Pisa, D. Alves. *Data Access Service in a Computational Grid Platform Applied to the Monitoring and Control of Epidemics on Georeferenced Dynamic Maps.* Proceedings of the 2nd IEEE Conference on e-Science and Grid Computing, 2006.

[12] The Globus Alliance. http://www.globus.org. Accessed in February 2007.