Integrated Surveillance Seminar Series
The Public Health Grid (PHGrid): Overview and Value Proposition

Tom Savel, MD
Medical Officer
Director – NCPHI Research & Development Lab
National Center for Public Health Informatics, CDC
March 26, 3009
Overview

- Public Health Challenges
- What is Grid
- Value of Grid to Public Health
- Current Activities
- Achievements
- Looking Ahead
- Research Activities
- Q&A
Current Challenges in Public Health

- Public health data widely distributed
- Volume of public health data growing rapidly
- Many cultural, social and political impediments to data sharing
- Requires a stronger economic model for long-term financial sustainability
- Uniquely dynamic, complex and global in scale
  - Situational awareness, population health, event detection, inventory tracking, countermeasures administration, alerting, etc
- Many redundant systems, application silos and data silos.
Current State of Public Health Surveillance (Data Flow)

- Intensive data gathering from medical facilities, state & locals into a giant CDC owned data warehouse
- Heavy use of statistical algorithms to detect anomalies in the data and trigger investigations
- CDC Centric Approach to developing and deploying software

CDC

State, County, and Local Health Departments

Laboratories

Practitioners
Current Challenges

- Politics of control of data has been the primary obstacle to formation of a national system
- Much existing data remains siloed at the Local/State level – accessibility and visualization limited
- Building systems non-collaboratively leads to low adoption rates
Model Formulation: Health Protection Framework

Outbreak Management System Countermeasure Tracking

PHIN Messaging PHIN Vocabulary

Supply chain management

Surveillance systems (NEDSS) BioSense GIS

Actions

Decisions

Knowledge

Simulator

Biointelligence Center Epi-X

Data

Information
Health Protection Framework: Foundation

Actions | Data | Information | Knowledge | HEALTH THREAT

Plans | Decisions

EDUCATION
POLICY
EVALUATION
RESEARCH
APPLICATIONS & SERVICES
STANDARDS
Illustration of Integrated Solution

Practice Alert in EHR prompts changes in physicians’ behaviors

Alert is embedded into Electronic Health Records (HER)

NYCDOH issues alert on Legionella Outbreak

NYCDOH decides there is a Legionella Outbreak

Index case of Legionella reported

Surveillance system @ NYC Department of Health detect cases of Legionella in Parkchester community in Bronx

Data

Information

Knowledge

Decisions

Plans

Actions

Legionella

NYCDOH uses alert on Legionella Outbreak
Action: Influencing Health Provider Behavior
Future Goal: Federated Architecture (Grid)

- **Leverage Existing Capacity**
  - Distribute resources and infrastructure
  - Increase flexibility and scalability

- **Provide Local Control** of data and services
  - Reduces political barriers
  - Address many privacy concerns

- **Foster Collaboration** to define requirements, priorities, develop, and deploy technology

---

**Local/State Health Dept**
- Surveillance & Informatics Capacity
  - Local/State Data

**Users / Experts**
- Scientific and Public Health Priorities

**Public Health Grid**

**Analysis / Visualization Capacity**
- Local/State Data

**Academic / Industry Partners**

**CDC & Other Federal Agencies**
- National Data
- Standards, Services, Guidance
What is a Grid?

- A computing infrastructure
  - provides dependable, consistent, pervasive and inexpensive access to data and applications.
  - By pooling federated assets into a virtual system, a grid allows data owners to share data and applications while maintaining control.
Grid Represents…

• Different way of thinking
• Different way of solving problems
• A long-term, fiercely collaborative approach
• Large-scale Computer Trends
Types of Grids

- Computational Grids (virtual super-computer)
- Collaboration / Access Grids
- Data Grids
- Dynamic Combination
- All on same platform
Computational Grids

Most famous/infamous

A massively distributed computing environment composed of over 3 million Internet-connected computers launched in May 1999 – has led to a unique public involvement in science.

Three million PCs deliver 6,000 CPU years per day – the fastest (admittedly special-purpose) computer in the world
Grid application models protein folding & misfolding (1224 teraflops, as of 23 Sept 2007)

Grid application models the way malaria spreads in Africa and the potential impact that new anti-malarial drugs may have on the region

Grid application models the design of new anti-HIV drugs based on molecular structure (in silico)
Other Computational Grids

Shared resources at San Diego Supercomputer Center, Indiana University, Oak Ridge National Laboratory, National Center for Supercomputing Applications, Pittsburgh Supercomputing Center, Purdue University, Texas Advanced Computing Center, University of Chicago/Argonne National Laboratory, and the National Center for Atmospheric Research
Collaboration Grids

- Presentation, visualization and interactive environments
- Runs on the same grid as the computational grid
- These combined resources are used to support group-to-group interactions, large-scale distributed meetings, collaborative work sessions, seminars, lectures, tutorials, and training

Access Grid aids SARS patients

During the 2003 Sudden Acute Respiratory Syndrome (SARS) virus outbreak in Taiwan, Argonne’s Access Grid technology enabled radiologists from across the country to review patients’ X-rays without added risk of infection. Three members of Argonne’s Mathematics and Computer Science Division assisted in the fast deployment of the specialized Grid.

Access Grid technology, conceived and developed at Argonne, connects people across town or around the world to interact and exchange ideas with each other in real time through high-resolution video and voice over the Internet. Each “node” in the grid may be able to display a dozen or more real-time images of other users, documents, Web sites and other useful data.

Normally, laboratory and academic groups use the Access Grid for training and education, conferences and workshops, site
Data Grids

- **CaBIG** – Cancer Research Datagrid
- **GEON** – Geosciences Network DAtagrid
- **EGEE/CERN** – The world’s largest particle physics laboratory…where the web was born (LHC – The Large Hadron Collider, May 2008)
- **DataGrid** – EU funded resource of shared large-scale database
- **TeraGrid** - Shares resources at San Diego Supercomputer Center, Indiana University, Oak Ridge National Laboratory, National Center for Supercomputing Applications, Pittsburgh Supercomputing Center, Purdue University, Texas Advanced Computing Center, University of Chicago/Argonne National Laboratory, and the National Center for Atmospheric Research
Open Source Grid Software/Projects

- PRAGMA
- Pacific Rim Applications and Grid Middleware Assembly
- eGEE
- Enabling Grids for E-sciencE
- HealthGrid
- DutchGrid
- Large-scale Distributed Computing in The Netherlands
- the globus alliance
- NGrid
- Open source grid computing
- gLite
- Lightweight Middleware for Grid Computing
- Open Grid Forum
- Open Forum | Open Standards

SAFER • HEALTHIER • PEOPLE™
Commercial Grid Consumers

Amazon Elastic Compute Cloud
Amazon Simple Storage Service

Pacific Northwest National Laboratory
Operated by Battelle for the U.S. Department of Energy
Gartner Hype Curve

Figure 1. Hype Cycle for Emerging Technologies, 2006

2-5 Years Mainstream Adoption

Technology Trigger | Peak of Inflated Expectations | Trough of Disillusionment | Slope of Enlightenment | Plateau of Productivity

Years to mainstream adoption:
- ○ less than 2 years
- ● 2 to 5 years
- ▲ 5 to 10 years
- △ more than 10 years
- × before plateau

Source: Gartner (July 2006)
Grid as a supporting technical framework for public health...

- Lets experts retain stewardship of information
- Flexibility allows Integration, Interoperability & Data Access between Silos
- In the long-term the Cost & Time to Re-engineer Existing App Silos falls
Value of Grid to Public Health – Meeting those Challenges

Ideal Attributes of a Public Health Grid:

• Open-architecture
• Federated
• Scalable
• Flexible
• Redundant
• Leveraging best practices

Thus, meeting the financial, social, technology, and security challenges...
Partnering Guiding Principles

- Volunteerism
- Willingness (without funding)
- Capability
  - Public Health
  - Technical
  - Ambiguity
- Thought Leadership
  - Public Health Research
  - Grid, Open Source (when possible), SOA, Distributed Systems
Research Guiding Principles

- Sustainability
- Low barrier to entry
  - Technically, financially, socially
- 100% Standards-based
- Reusability
- Collaborative
- Distributed/Federated
Current State of Grid Activities:
Research & Practice

• Methodology
  – First develop hypothesis and then perform research
  – Develop evidence base
  – Make evidence-based decisions on the value of potential tools & resources
  – Apply selected tools to both existing and novel systems
  – Move systems to production
  – Continuous evaluation and enhancement
Current State of Grid Activities: Research & Practice

• Leveraging:
  – Existing technology applied to a novel domain (public health)
  – An open / collaborative development process with our partners (academic, PH, industry)
  – CDC as participant – not has solution owner
  – An evidence base (not personal preference)

• Challenges
  – Gain expertise in the wide variety of grid-based resources currently available
  – Define the gaps between the PH and Grid domains
Results & Lessons Learned

Results

• PH informatics infrastructure is readily deployable in public health settings – over 10 nodes established

• PH data can remain where it is best secured – fusion biosurveillance data from different nodes without physically transferring data

• Simple PH Analytics / SA can be supported in distributed environment - results can be displayed in maps and epi curve.

Lessons Learned

• PH will likely be supported by multiple service providers

• Collaboration is key to driving requirements and resolving issues

• Weakest hardware or connection in a federated model can be the bottleneck for data visualization and analysis
Future Directions

• Move from Research to Pilot to Production
• Develop Community of Practice and engage more partners as nodes
• Explore security and interoperability between frameworks
• Features, Functions, and Priorities for Situational Awareness Services
  – Develop Ability to fuse and analyze data from heterogeneous data models
Final Thoughts

How should you think about grid.....
The Public Health Grid
Research Activities
Overview

- Tools
- Purpose / Mission / Objectives
- Research Activities
- Future
Context

15 Months
Tools

- **Sourceforge.net** for source code management (source code versioning and control)
- **Subversion** (Apache license) used by Sourceforge.net to store and manage the source code versions. Also used on our developer workstations.
- **Sourceforge.net** for issue tracking (bugs and feature requests) and product releases (service packages & tool packages)
- **Eclipse** for our integrated development environment (Eclipse license)
- **Maven** for our build and configuration environment (Apache license)
- **JBoss** for our portal application server (LGPL license)
- **Tomcat** for our grid node service runtime engine (Apache license)
- **Globus** toolkit for our grid node service container (Apache license)
- **Hibernate** for JDBC data access within AMDS (Aggregate Minimum Data Set) services (LGPL license)
- **caGrid** (NIH's Cancer Bioinformatics Grid) for service infrastructure (Apache license)
- **Collaboration Tools** (Google Blogspot, Sites [Wiki], SMS Texting, Instant Messaging)
Purpose

- Determine the viability of Federated Architecture in Public Health
- Establish relationships with key partners / collaborators
- Determine / Inform future public health informatics approaches
In view of improving the health of our nation and of our world through the practical use of innovative technologies, our goal is to identify, research and simplify computer technologies for use by both developers and users within public health practice. Core principles include: Long-Term Sustainability; Low Barrier to Entry (Technically, Financially & Socially); 100% Standards-Based; Reusability; Collaboration; Open Source; Best Practices; Distributed; Federated; and a Bottom-Up/Middle-out Approach.
Objectives

• Provide a secure, easy-to-use national technical and social infrastructure for solving public health problems
• Develop an extremely low cost grid appliance
• Simplify web services development (drag & drop)
• Simplify data access and data exchange (drag & drop)
• Connect public health grid to other grids, and to other data sources, regardless (in other words, interoperate with everything)
• Recruit local & state health departments, HIEs, RHIOs, academic institutions, national data sources, medical centers, international public health partners, and vendors
Current Proofs of Concepts
BioSurveillance POC: Federated Search

Goal: Explore standards based federated frameworks to promote distributed data stewardship, analytical access, and collaboration between participating stakeholders. Inform NCPHI and its public health and commercial partners of best practices and potential issues to this approach, and provide a foundation to evaluate existing and emerging interoperability protocols.

Primary requirements:

• Demonstrate the capability to share and visualize biosurveillance data:
  – Within a State
  – Between States
  – Between States and CDC
• Aggregate data under control of state, share results with external users
• Combine and visualize results in the form of maps and simple analysis (e.g. Epi Curve)
RODSA-DAI

Foundation: Real Time Outbreak Detection System and Globus Grid Toolkit

- RODS - ~20 production instances across US
- Globus – Leading Open Source Grid Middleware; used in NCI’s caBIG, GeonGrid

Hypothesis: Extending RODS with Globus Services allows the ability to query across installations, and visualize data from disparate / secured nodes
RODSA-DAI Demo

http://ncphi.phgrid.net:8080/rodsadai-web/
Poison Control Data Access & Integration

Goal: Research ability to augment public health situational awareness, by accessing non-clinical data sources of public health importance, based on secure web services

• Demonstrate access and visualization of poison control call data via web services
• Display data over multiple days over multiple call classifications
• Combine and visualize results in the form of maps and simple charts
Poison Control Demo

- PoiConDai

http://ncphi.phgrid.net:8080/poicondai-web/
Aggregate Minimum Data Set

Goal: Facilitate multi-state public health situational awareness with simple, common data interchange service based on a subset of key biosurveillance data elements

- Obtain consensus on most relevant elements
- Create common biosurveillance data structure aligning to AHIC / HITSP standards
- Develop interfaces to existing partner biosurveillance systems
- Distribute & refine using open source principles
- Proposed elements
  - Syndrome
  - Syndrome classifier
  - Patient 3-digit ZIP
  - Count
  - Date
Developing a Distributed Research Network (DRN)

DEcIDE centers at the HMO Research Network Center for Education and Research on Therapeutics and the University of Pennsylvania

Introduction

Background and significance

- The use, cost, and breadth of new medical technologies are growing rapidly
- Stakeholders seek emerging information about their relative risks and benefits
- Growing availability of routinely collected healthcare information
- Coordinated approach needed to generate evidence about the harms and benefits of therapies
Rationale

• To answer many public health questions, it is essential to use information from more than one electronic data system.
• Efficient ways are needed to securely access and use data from multiple organizations while respecting the regulatory, legal, proprietary, and privacy implications of this data use and access.
• Allow data owners to maintain confidentiality and physical control over data, while permitting authorized users to ask essential questions.
Project Goals and Objective

• “The primary goals … are to improve public knowledge about health outcomes in time frames that are quicker than traditional research approaches; and to take advantage of the power of networks…”

  -AHRQ DRN task order solicitation

• Objective: to design a scalable, secure, distributed health information network—a distributed research network—to conduct population-based studies of the risks and benefits of therapeutics
Current Project Activity: Proof-of-Principle Demonstration

• **Build a network proof-of-principle to demonstrate selected functions of a distributed research network**
  - An authorized user authenticates to a central portal based on digital certificates
  - A SAS program is distributed to each data owner (node); the data owner allows or denies the request for the program to run
  - The SAS program is configured based on the data owner’s (node) local SAS settings
  - The SAS program is executed at each node, and a standard results set is returned
  - The results are aggregated and made available to the authorized user
  - A log of site activity for each node is generated

• **Evaluate the proof-of-principle demonstration and characterize the needs, challenges, and barriers to creation of a distributed research network**
Proof-of-Principle Implementation

- Choice and selection of technologies for demonstration
- NCPHI’s role as partner with proof-of-principle implementation
- Overview of the development and implementation process with NCPHI, Informatics team, and participating sites
  - Geisinger Center for Clinical Studies
  - Group Health Center for Health Studies
  - Harvard Pilgrim Health Care
  - Kaiser Permanente Colorado
  - Kaiser Permanente Northern California
Technical Demo
DRN Lessons Learned

- Challenges and barriers to implementation of a distributed research network
- Suggested approach to development of a distributed research network
- Weekly coordination calls essential to collaborate with organizationally and geographically distributed partners
Service Registry

http://sites.google.com/site/phgrid/Home/service-registry
Future

- Move from Research to Pilot to Production
- Develop Community of Practice and engage more partners as nodes
- Explore interoperability between NHIN and Public Health Grid architectures
- Expand public health use cases
  - Build additional services
- C&A Globus and other services
- Public Health Node Appliance
  - Windows Version
  - Linux Version
  - Simplify, simplify, simplify
- Send node & services to data

New Emerging Partners
- CCID / Grid Computing / Pathogen Data
- Environmental Tracking
- Birth Defects
- Genomics / Bioinformatics
- NEDSS
- Emory University
- Georgia Tech
- Internet2
- ONC
- WHO (EA Lead)

Big Unknown
- Stimulus Package
Thank You!

Questions?