INTEGRATED SURVEILLANCE SEMINAR SERIES

CENTERS FOR DISEASE CONTROL

COORDINATING CENTER FOR HEALTH INFORMATION AND SERVICE

NATIONAL CENTER FOR PUBLIC HEALTH INFORMATICS

DIVISION OF INTEGRATED SURVEILLANCE SYSTEMS AND SERVICES

& COORDINATING OFFICE FOR TERRORISM PREPAREDNESS AND RESPONSE

TITLE: The Public Health Grid

SPEAKER: Tom Savel, MD (Lead, Grid Computing Research, CDC/CCHIS/NCPHI/OD)

MODERATOR: Mike St. Louis, MD (Associate Director for Science, CDC/COGH/OD)

April 26, 2009

RUTH JAJOSKY, EPIDEMIOLOGIST, DIVISION OF INTEGRATED SURVEILLANCE SYSTEMS AND SERVICES, CDC: The Integrated Surveillance Seminar Series. It's one o'clock so we're going to get started. I'm delighted to welcome you to this Integrated Surveillance Seminar Series. This presentation is being co-hosted by both the National Center for Public Health Informatics and COTPER – the Coordinating Office for Terrorism Preparedness and Emergency Response. Through this forum we strive to establish interactive discussion about the vision of integrated surveillance, identify issues and approaches to its achievement, and provide a mechanism for discussing best practices in the area of integrated surveillance.

Following our presentation today, we're going to have a Q&A discussion period. For the benefit of our participants via Envision, IPTV, webinar, and phone line, please use the microphone when asking questions. Microphones are located in the middle of the room and on the tables in front of you. If you're using a table

microphone, please make sure you un-mute the microphone before speaking so others can hear you. For those of you on the phone, please put your phones on mute during the presentation and do not place your phone on hold. If you place your phone on hold, we often hear music streaming through your phones and that's a little bit disruptive. If you need to break away from the seminar, please hang up and reconnect at a later time.

We are please to have Dr. Mike St. Louis moderating our session today. Dr. Mike St. Louis is a science officer for Global Health at CDC in the Coordinating Office for Global Health or COGH. He joined the CDC's Epidemic Intelligence Service in 1985 with the Enteric Diseases Branch in the former NCID - National Center for Infectious Diseases. In subsequent years, he worked principally on surveillance and research of HIV/AIDS and sexually transmitted diseases both in domestic programs and internationally. He served as the founding director for the CDC's global AIDS program in 2000 and then established and led CDC's country office and global AIDS program in Zimbabwe. He's been with COGH since 2005 working on a range of cross-cutting global health issues including: global surveillance, health information systems, and harmonization of laboratory strengthening efforts across CDC programs. He serves as the chair of CDC's Surveillance Science Advisory Group or SurvSAG. Dr. St. Louis holds a BA and MD degrees from Harvard College and Harvard Medical School and was trained in pediatrics at Massachusetts General Hospital. So please join me in welcoming Dr. Mike St. Louis.

ST. LOUIS: Well, thank you very much Ruth. Let me reach out and congratulate Ruth and everyone associated with this [Integrated] Surveillance Seminar Series for putting this on. I particularly welcome it on behalf of the Surveillance Science Advisory Group which really appreciates the multi-disciplinary efforts and activities and forums that you put together with this. As Ruth's kind introduction suggested, I've been around CDC for a long time and working in a number of capacities, but a lot on surveillance. And I noted a couple of years ago - I think it was two years or so now - that I started hearing this word, "grid." A couple of reactions: One, it seemed like a big idea – and, it's always fun and interesting and good to hear big ideas being bandied about, thinking about how we think about the future. Second, it is just tremendously resonant, I think, as a metaphor about just being able to plug in to an electric grid and being able to use that. So I for one am eager to

hear the full presentation and hear how this is going to apply to public health and offer us advantages in public health and surveillance in particular.

I think that as almost a little bit of a challenge to Tom as he gets into this – I would note – that when big ideas come along, they offer the opportunity to transform the field, but also the challenge with big ideas, with things that represent major transformation, is that it can be hard to get from here to there, to a major transformation, and sustain the momentum and energy to get there, especially if there are people experiencing complexities with that along the way. I think that simultaneously keeping the big vision of what that idea represents, but also being able to document to the participants that there is progress - there's incremental progress at the same time – is really an important issue. As Tom comes up, I hope he will be in a position particularly to address those issues of both the long term vision and the incremental steps that are achievable, that are going to keep the sense of progress as we head toward the big vision.

About Tom, he's currently the lead for Grid Computing Research in the Director of Informatics Research and Development laboratory for NCPHI. Tom came to the CDC in 2003 as a Public Health Informatics Fellow. After that, he worked as the lead for informatics for the National Center for Birth Defects and Developmental Disabilities. Following that, he came to NCPHI. At NCPHI, he has served as the Acting Associate Director for Science until this past year, when he shipped it over to start leading this initiative with respect to this Research and Development Laboratory. He did his undergraduate degree at Wesleyan, trained in medicine at Mount Sinai, [and] has clinical training in general surgery and family practice. He's been a technology columnist, I learned in the process of this, for the American medical news, and has worked in a variety of news outlets. So with the combination of medical, informatics and communication skills, we look forward to Tom's presentation on this important topic. Thanks, Tom.

SAVEL: Can you hear me okay? It's a real privilege to get a chance to talk to you about the stuff that we're working on. I think that it's incredibly cool personally, and I'm very excited in terms of the potential transformative effect it can have. And, I completely agree with what Mike has said in terms of incremental steps that are achievable and big ideas [that] are wonderful, but if it's just dreaming, that doesn't help us right now.

So we need practical solutions, but still [need to] think really big. With that, I'm going to talk about what most of you are familiar with, which [are] the classical [questions]: What are public health challenges? What are we having to work with? What's our environment? I'll then talk about, "What is Grid." Some people say it's this and it's that. Which do you think about when you hear Grid? We'll then talk about – given what public health is, which is of course is massive – what is the value of grid to public health? We'll then talk about currently what we're doing, what we're trying to achieve, what are next steps are. In a sense, we'll show you demos of what we're working on. And, then we'll take questions.

Public health has a lot of challenges. Obviously, one of the most tremendous challenges is that it's a large entity or enterprise that is dealing with massive amounts of data to achieve many different things. Surveillance and all these various efforts, whether it's GIS or…. you look at sort of the activities of public health. It is truly massive, and the volume of data continues to grow. We're not dealing with small amount of data; we're dealing with massive amounts of data, and now we're dealing with – potentially – data from electronic medical record systems. So it's only going to get larger. There [are] many challenges around jurisdictions and around political issues in terms of wanting to share information and services and data. So there are a lot of challenges. And, obviously cost is always an issue. We're not rolling in money, and therefore, we have to think about what is a very cost effective way we can do this. I have here: "uniquely dynamic, complex, and global in scale." Again, our domain is about as big as it can get. Not that it was done on purpose, but historically there's this term. They call it a "silo." Now we see and say, "Oh, look it's a silo." But at the time, it met the need. So we're dealing with some historical systems. And, how do we work with new systems as well as the old systems?

Sort of the classic way you can think about how things have worked is you have data from practitioners and laboratories going and local up to county and then to state and up to CDC. There's this sort of this conceptually, a vertical transmission in a sense. If you look at this example, the sense of - you've got various facilities in the DOD and the VA. And, trying to get this data, moving it along to the state health department where they can do some efforts, and then moving it finally up to the CDC – this one example just shows you classically how the activities in the data flows in BioSense had worked. But, there were a lot of challenges

around this because when you start thinking about data – first thing is there is security, there is privacy, there is control – these issues have to come up. But the goal is, and you'll see shortly, can we create an environment where those controls aren't seen as such impediments?

So I'm going to show you this, just to give you another perspective on all this, that when you think about an event or a health issue - whether it's diabetes, I think health threat is very general - it's about thinking about let's get some data elements [to] try to turn that into some information that's usable. So it's the number five versus five degrees. Or, trying to get more and more richness until it's knowledge – meaning, you can start to make decisions on this information. That a specific population is having a certain event that is occurring so you make a decision and you have a plan and then an action. What I'm showing you out here is classically some of the systems that have existed within CDC to attempt to address some of these issues. But the kicker is, if you flip it on its side – that object - there's a lot that actually has to go on underneath. And this is just to give you something to think about, that in order to achieve all of this, that you have to have education, [that] you have to help people understand that this exists, that there has to be policies set-up that can work with all this, that there has to be an evaluation site that says, "Is this system doing what we think its going to be doing?" Research has to go on to continue to push the envelope. There has to be – obviously as you saw in that diagram – specific applications or services to meet those needs. Everyone has one, but you have to have a standard. Therefore you can have cross-communication. I'll explain in a minute. In this one example, what it's showing is the same diagram saying, "Okay, so if we have an electronic medical record, and you start looking through this various pathway from a case reported of *Legionella* to then having information about it and then making a decision about this information would ultimately – I love how it has to convert EHR to her, 'cause that's the Microsoft way – that the data can then go back into the electronic medical record." In other words, serve public health, direct clinical care integration – so this cycle as an example. Therefore – and you can see on the bottom – if you can actually, from the knowledge gained from public health, send a notification back to the provider, you essentially turn five clinicians into five astute clinicians. As a clinician, if you just say, "Heads up, there's something going on," I [as a clinician] already sort of have my antenna up ready for that and that makes a difference.

The future that we're trying to shoot for looks a little different. It's about trying to make all the resources that are around, whether it's local, whether it's state, or whether it's federal – that you can facilitate the sharing of not only the data, but the resources that are working on the data. And, that it's not about shipping all the data. It's a different model, it's more of, "Here's the data that I will allow you to look at. I'm going to hold on to my data, but I'm going to let you take a look at it." So in a sense, this different model allows for greater collaboration, less issues around, "Well, if I'm going to send you all my data, now you have to curate all the data that I was curating." Now there are issues around the control of data. It's looking at much more of a federated model, and it's not about all the data being sent to one location. So start just to think about, "Okay, he's saying it's a "non-centralized model."

So given that as your framework to think about, we're now going to give your quick update, an overview of what is this whole grid thing. Sort of the classic way to think of it is a *way* to provide, dependably, data and what we call services, and I'll explain those in a large non-centralized fashion. If one entity over here has data, and this entity has data over here, they are able to both leverage each other's data in a very controlled fashion. By pulling all these resources – as you can see here – it allows data owners to share data and applications while maintaining control. Now this explanation right now is a little more technical, but really it's also a social – and I'll explain that in a second. The key thing is it's a different way of thinking. It really is. You have to start saying, "Okay, it's not about everything coming to one location; it's about a different kind of environment." You can solve problems in a very different way than you could before. In a sense you don't even care that…I think it's 100,000 servers are being hit with every Google search. You don't want to know about that. But that's how they chose. And you would never say we're going to have 100,000 machines to solve a problem every five seconds. It's just a very different way of thinking. This is not an instant solution. This is much more of an organic solution. Again, just as with the internet, there wasn't a gazillion websites [when] it started off, and the websites grew as the environment was conducive for that. This whole thing is in alignment with what's happening now with regards to large scale computational trends. We'll go over that in a second.

So there isn't "A Grid." This isn't about "A Grid." There are different types of grids. The key thing to think about is, classically, there is a computational grid. We'll talk about that, but that's about computational

power, about a super computer. There's also collaboration and access grids which are about communication. There are data grids – which is what we are going to talk about a lot today – which is about sharing and exchanging data. And, of course, you can have combinations of these.

Briefly with regards to computational [grids], it's sort of been done a lot, but one of the classic ones is SETI@home. [It] was a screen saver that allowed you to take a piece of data - for "is there extraterrestrial life" – and do some processing on your computer, and then it would automatically ship back. So they were basically able to do [work] in a matter of a short period of time which would have [otherwise] taken enumerable years. Just to show you with the number of processing cycles that you can achieve with that - it says that three million PCs were able to do this in a very short period of time.

The [SETI] model was very successful so they said, "Why don't we try dealing with protein folding and other very computationally intensive work in a distributed fashion?" Down at the bottom you can see here the National Science Foundation, Google, Apple, Dell, Intel and the NIH are all continuing to move in this model. Other terms as you are out there learning about this – there's Biomedical Informatics Research Network and there's TeraGrid. So we are not inventing the wheel. The hat I want to wear when on my team is: We're not going to reinvent the wheel. Our goal is to find all the other wheels people have built and learn from those wheels and leverage those wheels. So we're all about leveraging what's already been done, but in a different framework, in a new way.

This just shows you the screen on collaboration grids. Collaboration grids were used with a lot of the SARS activity that was going on for communication very rapidly. So it was actually used real time.

The data grids – just to give you sense and one of the ones that we have spent a lot of time working with – [are] the Cancer Biomedical Informatics Grid effort at NIH. There's also TeraGrid, there's GEON. There will be many grids, and there are many grids because they have specific domains that they are working on, whether it's Geospatial or whether it's cancer research. But grids can talk to each other. It's got cross-grid communication.

So commercially, IBM, Microsoft, hp, Sun, Oracle – I'm shocked, of course – they have grid solutions. I have Tier 2 here meaning there's a great concept. This uses gigaspaces. When I'm talking about thinking

differently, this is an example. Gigaspaces - essentially they take a piece of information and shred it into pieces and scatter it over the grid and then when you need that information again it recombines it. So ultimate security [is achieved] by essentially leveraging the grids distributive fashion to secure information.

Now one of the key things that we are focusing on, given the specific challenges within public health, is the open source community. Saying, “Okay, we don’t want to tie ourselves to a vendor, so is there in the open source world a lot of activity going on with regards to distributive computing and grid?” And [for] those who aren’t familiar with open source, it is the concept where the code is available for you to use, make changes. There is, of course, governance and coordination; you can’t send it back and all these issues. It’s about not being tied to a vendor solution. Just to show you a couple of examples here, there’s Pragma, EGEE, Health Grid, there’s gLite. And you see up in the upper right, there’s the Globus Alliance. That’s who we’ve been working closely, with and I’ll talk about that in a moment. There are always tradeoffs. One of the reasons that we have aligned ourselves closer to Globus then others is because no solution is perfect. We have a lot of additional partnerships that we can leverage. Part of that is the solution itself as well as the knowledge we can gain from partnerships. So there was a compromise. We’re getting a lot of knowledge from our partners.

Commercially, who uses grid? Of course, Google does, Second Life does, Amazon has their Elastic Cloud, Boeing, Goodyear, Partners. So as you can see, we didn’t just think of this. It’s us saying, “Okay, there’s something they know that we didn’t realize that we should be looking at.” As I continue to give this presentation, my Gartner Hype Curve starts to get older and older. This is from 2006, and it’s been now a couple of years, but at the time they made this, grid computing was right in between the peak of inflated expectations, of course, and the trough of disillusionment. But now as you can see it’s probably moving over to the slope of enlightenment. Meaning grid doesn’t solve everything. Okay, maybe it does. No, grid doesn’t solve everything, but it can do a lot.

They way I try to think about it is: Grid is a supporting technical framework for public health with the key point that the experts don’t have to start saying, “Here’s all our data.” No, they can still retain control because, in many cases, those [that] are the stewards of the data are the best to be the stewards of that data; they know it well, they can maintain it well, and you’re leveraging their strengths. It really facilitates integration,

and, in time, reduces a lot of the cost. So you can see here it's sort of the picture of Epi-X and BioSense that it's about really creating a common connecting framework. Ideally, if we were to create an environment that was based on this distributed computing for public health, what would it look like? Well, it sounds a bit like mom and apple pie, but it continues to remind us that it needs to have a sort of open-architecture: you can see how it's built, there's no sort of behind the scenes magic, that it's federated, that it's scalable, that it can really grow in time, it's not going all of a sudden collapse on it's own weight, that there's a lot of redundancy, [and] that the term node if for example if one node goes down there's another one to take its place. Of course, we're going to try to leverage the best practices wherever we can. When we get into the question section, people say to me, "Well this sounds just like the internet." We'll talk about the difference. Therefore, we believe that this can really meet many of those financial, social, technological, and security challenges.

This diagram we've used a lot, but the key take home from this diagram is that no one is in the center. Yes, there [are] people, there [are] consumers, there [are] federal agencies, there [are] health information exchanges, there [are] providers. I'm not saying in this case that there are physically big nodes and little nodes – just showing that the grid is agnostic to that. So whatever entity says, "I want to be on this grid and I want to be able to work with other entities," they can do that. As you can see on it, it's about sharing. It's about sharing surf services and security services and aggregate summary data services and whatever public health wants. With the partners we've been working with, with regard to our research, we said, "Oh, first of all, we just need a volunteer. I hope that's okay." Then we need you to have knowledge in either technology, or public health. Or, if you're just good at confusing things for a while, [it is] because we have a lot of ambiguity. In a sense, what we're doing is we're crystallizing what others have done and re-digesting it and saying, "Okay, here's what they did, how does that apply to us?" Then, of course, a lot of people who have a lot of expertise – It really hasn't been about us saying we know what to do, it's what are the other experts in this domain – what are they saying? We met with some of the leaders of grid, and they said they felt that public health had one of the best use cases they've ever seen for distributive computing, and they explained why, and it really starts to make sense.

In terms of our principles as we move forward we said look, "This has to be sustainable. This can't be so heavy that no one can afford to do this. It has to be really sustainable. It has to be easy. We want to make this as

easy as a Firefox install." That has to be our goal. That has to be standard-based. That uses all the standard web services – it uses standards for security. Standards. Standards. Standards. This is not about building a one-off. Everything needs to be reusable. If someone builds a component that is amazing at doing CUSUM analytics, another entity says, "Hey, can we just use that? Yes, it's available. Here's the information you need, and you can use it right away." Of course, it's distributed or federated whatever term you like to use.

We are trying to follow standard mom and pop apple pie of: develop a hypothesis, develop an evidence base. We're trying to make decisions saying, "Well, I just like this technology because I think it's cool" - that doesn't fly with us. We're saying, "Okay, I think it's cool, but now can you prove why you think it's cool. Give me some evidence. Show me why this is the way we should go because there's many different options." Then we select some tools, we move things to production, and then our goal is to continue to evaluate and enhance what we've decided on doing. Again, we're trying to leverage this open collaborative process with our partners, whether they're public health partners or academic. CDC is a participant; we are not the owner of this. Essentially, we want to act like we got the ball rolling, and now let's run with everyone else. We are essentially a very enthusiastic participant. Yes, of course, we nudge the ball a little bit at times when we have to, but really, it's not about us. It's about the whole community and what they need.

We've had some challenges of course because we went from "zero to hero" of understanding, "What does Grid do? Where's the hype? What [is] the reality of what grid can do, and what it can't do?" So we've had to, sort of, in many cases – we converted one engineer into a grid engineer. And now we know how to do that because this gets into the big issue around the public health work force. Your like, "Oh, great! Now we're going to need grid people, and how are we going to get them out into the field?" The goal is to make that barrier very, very low. That's our challenge; to get this at a point where it's not that much of a jump to get us there. Of course, define the gap between what grid provides and what public health needs and really work on that.

So what did we learn so far? I'll go into some details. We learned that this is absolutely doable; that it's not just hype. We'll talk about this. There's this concept called a node, which essentially is a creature or an entity on this Grid. Think of it as a website or a server. I'm being very blunt about that, but a node, an entity, [is] a point that wants to be able to share as a user or a consumer. That we can actually – and I'll show you these

demos – do this in a non-centralized fashion. That for the first time, instead of saying, “Send us your data and let us do the analytics.” You leave your data here and we’re going to be able to actually get the answer we need without you sending us your data. And collaboration is critical.

So what I’m going to show you now is our future direction. Our goal of course is to move from research and pilots to actual production and working on really enhancing this Community of Practice - and I’ll show you how we’re doing that - and making sure that this is as secure as anything else that’s in existence and understanding, “What are the priorities for the various functions that we need?” We have to prioritize. What are we going to need? What’s sort of the killer app for the Grid in the short term?

So before I start showing you a few more things – how should you think about Grid? I think you should think of it as both a technology infrastructure as well as a social layer for collaboration that is not centralized and has the ability to scale because it is not centralized.

Let me give you a sense of some of our activities. I’ll give you a bit of a sense of what are the tools that we’ve been using and what are our objectives and then some research activities. Everything you’ve seen we’ve done in 15 months. I’m very proud because I’ve got a wonderful team to work with, incredibly bright, hardworking, and I have to say, I learned quite a bit.

This is one of those pages where (missing) would not be very happy since it’s just a huge amount of text. But that’s fine. It just gives you a sense if you just look at some of those words. Some of them may ring a bell, some of them won’t. But, we are practicing what we are preaching. All the tools that we are using are based on open source. All the tools that we are using for our infrastructure, for developing the code, for sharing the code, for all these various aspects, are all open source. So we are practicing what we are preaching. And I’d be happy later on to talk about all these things. I didn’t understand half of them, although I’m an informatition. I actually didn’t understand half of these before we started. It’s really impressive how these all fit together.

The purpose of these is really to show: here’s what we think it can do, can it really? Again, this is everything I briefly mentioned, that our Mission Statement is all about low barrier to entry, sustainability, open source, [and] collaboration.

This just gives you a little more detail the objectives. For those who really like technology, this is really about simplifying everything that we could. This did not take someone who was a brilliant coder to get this to the point where it's a drag-and-drop. Because if you don't make it that easy, my perspective is that public health is just not going to use it. They've had enough of this.

Way back we called it the Biosurveillance Proof of Concept. And what this got at was if you think about the goal of Biosurveillance, which is trying to, at a very high level, get a sense of what's going on so you can make a decision. Could we do this in a non-centralized fashion, aggregating data from multiple users and then being able to visualize those results? We ended up working with the RODS team which stood for the Real Time Outbreak Detection group, and we worked with one of the groups in the United Kingdom to figure out a way, in a sense, to "gridafy" databases. Grid – classically, their strength was about computational power, but could we "gridafy" a database? In other words, could multiple databases that are around the country be seen as one database?

I'm going to see if I can do this correctly. So this is the live demo. If you take a look, again this is reusing, not reinventing anything. Let me just show you right down here. So what we said is, "Look. Let's take Google maps, [and] let's have - using the grid - let's have two databases that have grid nodes." So in other words, take a database put a grid node there and expose that database at the grid node. One of them would be located at the CDC and one would be located in Dallas, Texas. Now, and I'll do it for you live. It's a very primitive demo, but it shows that you can choose to – you can select one or the other. And so in this case we'll just pick Dallas, and I'm going to submit so it's going to go to the Dallas node [to] pull back the data from Dallas. In this diagram, these are counts of per, in this, zip code, how many counts of fever there [are]. Essentially, how many individuals have fever in this location? You can see the different colors represent the counts. White: 1-5, green: 6-10, yellow: 11-15 and so on. So if we said, "No, let's combine both different nodes,"…and now we'll submit. It's going out to both nodes in the grid, getting that database, and sending back just the summary count data, combining that data back here. And now, if you actually look, you'll see there's more red because it's combining counts. This would be your classic example of two different hospitals systems

that have an attachment area and can get a sense of what's going on without making them ship us all their data. This was using RODS data to do this. So that was the one example I wanted to show you.

So next we took a different perspective. We said, "Okay, well first of all, this is really about not sending the data everywhere. So what if instead of sending data, we let the entity that has the data hold onto the data and we at CDC will essentially ping their data and just pull back a subset summary of that data?" This was not about shipping data for those techies. This is about using web service to pull data, but allowing those areas that want to hold on to their data, hold on to their data.

The ability to augment public health situational awareness by accessing non-clinical data sources – so, in this case we're looking at poison center data, we're looking at how many reports of nausea were there on a certain day at a certain location.

So here you have an example, again, little different than before, but using Google Maps. What this does is essentially takes - at CDC - just pulls their data – if you think of it as their data is five inches thick, we just said give us the first inch, just give us a piece of that data that will let us get a sense of the various – and you click up above…and this says "I want to look at all clinical effects". Meaning I want to look at everything, not just nausea, or vomiting, or diarrhea. I want to look at all clinical effects within a certain date range in all the counties in this area. And this worked, so we said okay, great. So we can do a distributed and we can then pull just a piece of that data that we need. So that worked.

And we're working on another one. Which again is the same concept, but we're sort of refining the wheel each time we take a pass around. And we're working on something called, for example, the Aggregate Minimum Data Set. This is saying, "Could we have multi-state public health situational awareness with the simple common data interchange service based on a subset of key Biosurveillance data elements?" Meaning, this gets away from the concept of just because you can pull and transfer tons of data doesn't mean you should. So this whole philosophy said, "let's flip it on its head." If you only had five data elements, what would you pick? This is really trying to get at: can we get a small subset of data from multiple nodes? This is really trying to say: just give me the first quarter of an inch of all this deep data. This is what we've been working on. So you can see we're just moving along saying, "If all of a sudden we can get syndrome – how ever you want to call it

– a bidding of some diagnosis." The classifier is saying, "Well, this is from BioSense and this is from Essence," and you're able to say, "Where was this patient and how many on a certain date through the entire country" – the value in that. So it's a different way of – in a sense it's going retro, back again saying, "Yeah, we're in a time now of tremendous amount of data, but can we do this with just minimal amounts of data." So it's a different way of thinking.

What I want to show you now, real quick, is this. I want to change gears just to break it up a little bit. What I want to show you is this, which said we need to have a way to share and communicate with are partners about all this exciting stuff going on. As you can already get a sense, there's a lot going on and we said, "You know, emails are not going to cut it," and so, immediately we created a blog. And that blog, I think already – there [are] already four or five hundred posts. And everyday, when activity is going on, we are posting information about this. So this blog is extremely actively used to share information about what's going on. And as you can see on this, there's lot information and discussions, all our lab projects, who are the participants – as you can see that list goes on and on. If you want to get access to the code, when we've had things in the news, web resources, documents – we needed a place because of all the information going on. This is now saying, "How are we managing this community that's really very active?" And that's with the blog. But then it turns out the blog wasn't enough for us. And that's when we created a wiki.

This wiki allows us to put on documents, and issues, and use cases, and diagrams, and a place to go to start finding out about what everyone is thinking. And one of the pieces that I'm very proud of, it's called the service registry. In other words, if you think of what I've been showing you, as we call them services – think of them as tinker toy components that fit into the grid. So the grid is the framework and within that we create, anyone can create a service that's based on standards and [anyone] can use it on the grid.

As you can see here, for example, we have a lot of our partners, and one thing we are really proud of – if you haven't heard of this – one example, it's called Medley. It's a natural language processing engine. It takes text blogs and digests it and produces data behind it. In other words, in this discharge summary, there were issues around a patient who had diabetes and then had an MI and then had other issues – Medley would take those and find those words and put a control vocabulary concept behind those terms and [it] therefore becomes

machine readable. One of our partners worked on “gridifying” Medley. In other words, they have Medley, they called it Grid Medley, on a node and that you can be anywhere else on the Grid and say, “I have a text blog,” and you could just send it to Grid Medley, [and] then get back your results. So we’re very happy that, as you can see this list – and you can all go here – and you can look at the wide variety. What are we up to now, 25? 27. This is the equivalent to me of webpages. These are tools and resources, now that someone’s like, “I would like to do some visualization. I want to do some analytics.” You would go here – and this is just the first generation – and then you could literally, you could go to this and find out what’s the web address and how do I get there, and how do I plug this in and use this? Really, it’s about creating all these components that are behind the scenes so you can now have a lot more functionality than before. Just like with Google, you don’t care what servers they are hitting. That’s what this is about. It’s continuing to grow. This is first generation. What the master plan is going to be in time – because this is grid computing – many if these services in time – just like when you have some quad core processer that you’re using Microsoft Word on, that in time you can use your quad core processor to do incredible amounts of work – we’re going to be creating these next generation services that can do processing in minutes that would take hours. Essentially this is a Ferrari and we are just learning to drive it slowly, but there’s a lot of potential. So I wanted to briefly show you that service registry.

What I want to show you next is this. So you saw that first diagram that showed you the poison data. If you take a look now you can say, “Well, that looks a lot different than before.” And what we’re working on is creating a reusable component that works with, for example, just Google Maps, that whatever you want to visualize, you can visualize with this way with count data. And then to take it to the next step; if you actually mouse over and click on one of those areas, you now have actual data to look at. You actually can look at real time – I mean in terms of, whenever you click on this link you can start to customize this. The interface is beautiful on this. You can actually look and say, “Okay, so the case count on this date was six…and the running total…here’s the average.” So literally you just very quickly go to the map, click on this area and say, “Ah, I can see what’s going on” – making this completely reusable for whatever kind of data. In other words, it’s about data match ups. And that’s where it’s going.

I'm going to briefly go over this. What this was about – so CDC loves SAS. What this looked at was saying, ""Okay, with regards to a Distributed Research Network that's actually using clinical data sources, where normally they're going to be using SAS, and they take the SAS query and they have to send it to each one of their partners, and each partner works on it and then sends them all this back – could we "gridify" this SAS query?" And the answer was "yes." So in other words, you have your SAS file and what you want to have done and you essentially have the other grid node – are waiting for this to happen – receive that SAS query, process it and then send it back. So in other words, we've now distributed SAS queries. It's a whole different way of doing business. If you take away from this saying, "Wow, you really can do things differently." That's the great take home message that we can start to do things differently just like with the iPhone. You can start doing things and having a touch interface that you never had before. It gives almost public health a new way to think about how to do things.

Again this just goes over some of the steps we went through. We worked with AHRQ on this – The Agency for Healthcare Research and Quality. Because their challenge was, part of their task was - can this actually take place? So this is what I just described. We worked with our partners Geisinger, Harvard,[and] Kaiser Permanente. This is your techie diagram that basically shows that we had lots of security and firewalls, and that this can be done. I showed you the service of registry. This is really discussing the same thing that I just briefly mentioned prior, it's about working with our partners inside CDC and outside CDC. So within CDC we're actually working with CCID and looking at how to actually integrate this with bridge computing and pathogen data, environmental public health tracking, birth defects, genomics, [and] bioinformatics. It's a perfect fit when you start dealing with massive amounts of data. If you said, "What's the environment that can handle – can handle if needed – massive amount[s] of data?" It's a grid-like environment. And we're working with Emory. Emory has a lot of experience with grid computing. A lot of the work we've done, we have been leveraging expertise that caBIG, Cancer Bioinformatics Grid, has done. Because, again, why reinvent the wheel? So they're really becoming a great ally for us: Georgia Tech [and] all of a sudden now our National Coordinator, [be]cause clearly, this is about integration nationally. And, you can't talk about that without talking about NHIN and a lot of those issues. So we're absolutely making sure that whatever we're moving

forward with is completely integrated with the efforts going on in the Office of National Coordinator, the Federal Health Architecture activities, and the overall NHIN work. And from our perspective, we're all exactly moving in the exact same direction. We can talk about the relationship and how it will fit well together.

Some of our specific challenges for those who like the geeky stuff [are that] a lot of this has been working on Linux. So think of it as it turns an operating system into an application. So literally instead of double clicking on Microsoft Word, you would double click on this grid environment that would launch, and it essentially becomes virtually a virtual computer on your system and works. But we're working also with direct Windows installation. So again, our challenge is making this easier and easier to install.

So with that, I'm actually done a little bit early, but you know there's so much talk about, and I think it's more important to give you a flavor of this and of course come to our lab, talk to us, give us your perspective, join in. But it's been a pleasure talking to you today. So thank you very much.

ST. LOUIS: Well, Tom, thank you very much. It's been very stimulating. I'm sure there will be a lot of questions. There's a couple of questions, two questions, that won't be necessarily so quick, but will give time for people to collect their thoughts and prepare to ask some other questions. The first one is your reference to this new strategy, the Grid Strategy, and my question is: what do CDC scientists need to do to be in a position to work with this, support this, etc? You know, you take the example of Google, one of the inventors or first pioneers of grid computing/cloud computing, but we didn't know about that until later. All we knew was that we could search better, faster. And it was something we wanted to do and we had it there. Maybe that's a segway to the other point you raised to about a killer app or some early quick win that will let scientists across the agency realize that this is for them. Or, alternatively, how do we need to engage everyone to appreciate this? And the second question is simply a security one. Since of if we keep tightening security we're going to lose access to our flash drives, we're going to have to wear chips that we have to stick into machines and so on. But the idea of opening up and being integrated with the world raises questions then on how openness and security. So maybe you can just take on those first two things and then we'll switch over to the audience here on the multiple electronic media to answer some questions.

SAVEL: So we'll tackle security first. And I think the key way that I like to think about this, and I think the answer which will help make it clear is, well, what is the difference between this and the internet? The answer is this provides a layer of security which normally the internet really doesn't provide. So in other words – and I'm happy to talk to you personally one-on-one about this – it leverages what's known as those certificates of authentication, 509 certificates. The grid does not say you have to use security, but the grid does say if you would like security and you would like high level security, you can have that. So the answer is essentially, the Grid infrastructure is not the Wild West. It can be as secure as any other infrastructure where you have to have multiple layers of authentication authorization. So the grid can meet that. I think what's also valuable is it's less – you can ship your data, but the Grid doesn't require you to. Because, for those in technology, essentially it's a secure services oriented architecture; therefore, it's about distributed services. That's the other thing; which I see it as much more palatable because they don't have to start shipping all their data, but they're like, "Look, I will essentially customize the window into my data based on who you are on these five elements to start with." So in other words, you can start small and grow as you get comfortable.

With regard to CDC in terms of what should the scientist be thinking about, I would want them to say, "If what you're saying is true, let me send you an email about what we would love to see happen. Here's this data over here, here's this data over here. Is there some way we could leave those there and be able to combine them?" I would love people to say, "We have a need, but we never thought we could do this." Start telling us. We are moving forward with some of the ones we know of, but public health is massive. And so right now we are looking into, for example, laboratories and clinical labs not having to centralize all their data, but be[ing] able to actually share data while still holding onto their data – which is very palatable. I'm very excited about the impact I think it can have with regard to laboratory [data]. So that's my final perspective on those two questions.

ST. LOUIS: Okay, well if there are questions in the room. Henry if you could please come to the mike. And then we'll roll through our different types of electronic media to look for questions. Please introduce yourself.

HENRY ROLKA: I'm Henry Rolka from the office of Critical Information Integration and Exchange. Thank you very much, Tom. I know you've been pushing this forward and it looks like you're making progress. I would like to ask a question regarding a model, and the model is Vaccine Safety Data Link which for ten years, people worked together across multiple HMOs. It was sort of consolidated and coordinated through CDC. And people shipped data. It was on a yearly cycle. And they did studies on rare events needed a lot of data across a big geographic region like that. Since then, the participants have stopped sending the data – and I may be totally up on this, it's been a few years since I've been involved in it – but, they started writing SAS programs and distributing them so that they were run at the individual HMOs. And then the information results were consolidated just as you described, and you had a slide up there that actually listed a couple of those HMOs. In the development of that model, you had people working together for ten years and developing sort of the culture of data management and the understanding of the kinds of queries that would be involved and organizing the data in a framework and architecture that would be amendable to sharing like that. In other words, we needed to race; we needed a fast car, and eventually, over a period of ten years, built the Ferrari. Now you're sort of reversing the model: you've built the Ferrari and you're looking for the race. Have you talked about an implementation strategy or thought in terms of the CDC culture how you would go from having the tool to the programmatic needs and putting those together?

SAVEL: What I am very happy about is we went with a very light weight Ferrari and we made sure that there was nothing much in the Ferrari that we didn't need. So we basically went with – it's a one CD install essentially, and what it's really about now is – gradually, over time, as you saw from the registry sheet – this was what some of our Centers for Excellence felt would be very nice components to put in the Ferrari. In a sense, over time there are going to be more and more components in the Ferrari. Which means, over time – just like when you get a browser now, your experience is so tremendous versus when the internet was a little younger and there weren't that many web pages and people were scared of using Amazon. So I think it's a matter now of the grid [that it]doesn't require a certain number of nodes, it can work with two nodes. Like we

just got a call today from I think one of our partners in Utah and they said, “We would like to install one of your nodes and start to take a look at it and see what’s in it and what can we start to work with.” So I think it’s a matter of, at this early stage, as we start to build some of these core services, working with our partners and saying, “Okay, so you guys like SAS. So we’ll work on some core SAS services and put those into the Ferrari.” Or essentially, once you install the Ferrari, it’s going to say: “Do you want SAS services? Do you want these services? Do you want those services?” So it’s really going to allow people to customize and it’s very lightweight to begin with so they’ll be able to put in and I’m hoping to actually very soon to have a lab exchange service. So I hope I answered it, but basically lightweight engine which is completely flexible to put in one or a hundred services. The core Ferrari that we got – I think it has a hundred and something core services and we removed like a hundred of them, because we didn’t need them for public health. And we’ll slowly start to add back because right now our goal is [that] we want to keep it simple. We think that’s going to be much more tasty to public health saying , “Just give me what I need to get the job done,” and, “Oh, look! Now I can do this.”

JAJOSKY: We have a question on the webinar. The question is: Are there processes to evaluate data sources from a data quality perspective both prior to the data source joining the grid and on an ongoing basis?

SAVEL: So the answer to that is: not yet, but sort of by definition. You don’t have to be a grid node to use the grid. You have to be a grid node if you want to be a sharer or producer of data. In order to do that, there are standards with regard to data. Let me catch myself. There’s always going to be issues around semantic and semtatic interoperability. The data standards, the LOINC code, the SNOMED code. And the grid doesn’t sort of get rid of that. [The] grid is just saying we’re going to create a common framework which that standardized data needs to be exchanged. With regard to – for example, if you’re going to share let’s say summary laboratory data, that there will be a service that says here are the elements that you will be exchanging. Quality is always going to be a challenge and I think part of that will be, part of the effort of this community will be assuring the

quality and some validation services for example. Yes, this is on our radar of things we need to work on, so great question.

ST. LOUIS: We'll take one question here and then we'll try to go to Envision for the participants.

CONRAD HYASHI: Hi. Good afternoon. Conrad HYASHI. Division of Bioterrorism Preparedness and Response. I really salute the work that's being done here and I see some exciting potential applications. A question I've got that refers to how you use Glubus and things like SETI in supporting large scale linkages of our PC networks. A very practical one is, why do we not have at CDC, [have a] plug-in say which plug-in default search capacity would you like to be able to support today? Or, [have] a determination at the central level to say, "Hey, we've got all these computers, we should be able to be using them day to day". To me that's a waste of computing power, but I don't know if that's the case or not.

SAVEL: I can always say I hope in the future as I know of some other initiatives through [the] health grid that internationally they are actually using hospital computers on down time for computational power. I hope CDC, in time, will leverage grid and leverage its own computers when they're down for computations. I hope so. And I'll bring that up whenever I can.

ST. LOUIS; Let me ask now for the Envision audience. Those on Envision are on the telephone. If you want to ask a question please go ahead.

JAJOSKY: So we do have another question on the webinar as we wait for the folks on the telephone…

ANONYMOUS: Hello… Yes, I guess someone put us on hold. I want to talk about the vocabulary question a little bit more. We here at CDC have probably spent hundreds of man years sitting in meetings trying to harmonize vocabulary. I know, like, at the National Library of Medicine, they have like twenty seven full

time vocabularians whose job is merely to maintain the vocabulary list to write documents with. I wonder if that's not a bottleneck in the process. Maybe we need to have something where the site advertises our own vocabulary as opposed to agreeing on something in advance. I'm thinking particularly in terms of something like the STNX protocol where you advertise your vocabulary and what's available. You may have different nodes available at different privilege levels which get different layers of activity. So that's my question.

SAVEL: So that's a great question. There are many different ways to tackle vocabulary problem. The Grid provides a framework to tackle many of those different ways. There is always going to be some pain in transition. Anyone that says there is a quick fix is not, I would say, being realistic in terms of going from non-standardized to standardized [vocabulary]. It's going to be about creating services that either do translation mapping or that are able to do some processing to achieve these standards. There's going to be efforts around standards regardless, but if you end up getting a big bang for you buck that by getting to the grid, by taking the effort to standardize your data set. You're like, "Wow! I have access to specific subsets. It's definitely worth that effort. So I think that's really a matter of (inaudible) environmental data that you never had before that it becomes - because you make that transition. And once you make that transition, maintaining the quality of standards is always a challenge but is worth it for the benefit in the long run.

ST. LOUIS: Ruth, we have a webinar question?

JAJOSKY: Yes, we have another question on the webinar from Brian Fowler. And the question is: What level of redundancy will be needed for Grid technology to work efficiently and effectively?

SAVEL: It's a great question about redundancy. And the way you think about redundancy, is when I say you have to think differently, it's a whole different way of thinking about it. And I like to think of the grid as either sort of, whether you want to use eBay or Amazon, it's a very different model. Some services will be used a lot and some services won't be used a lot. And it turns out that certain services are getting incredible high

demand you put up another node and you replicate that service. So there will take a bit of overseeing. There's no question, you can't just set it up and walk away; there is a monitoring of: "Wow, this service is incredibly popular," [tying] into a whole different mechanism for funding, because you'd say, "Wow! If one service is being used a lot it should get potentially different amount of funding then those services that barely get used." Potentially, of course, there [are] some services that would get core funding because they create the breadth of services that are out there. So that's a great question of redundancy. I think there's no question that if there was ever an environment that could handle redundancy it's a grid distributed environment versus a secure.

ST. LOUIS: Alright, is there more questions from the telephone or Envision?

RICHARD DRAUGHT: Tom, could you talk a minute about the Community of Practice which envisions the next steps in moving forward? How it's going to be facilitated? Will it be apart of PHIN Communities of Practice and so on?

SAVEL: That's a great question. So as we speak we are in the process of transitioning and bringing the PHIN community into this community of practice. The reason we hadn't yet, is in a sense we wanted to have, got to have, this dinner somewhat cooked while the appetizers are ready. So we had some core services that we had the infrastructure fairly well set. We're going to continue to fix it and enhance it and tighten it and harden it. But we wanted to have it at least somewhere along. Prior to this, our partners were a lot of our grid partners, our Centers of Excellence, and some of our sort of leading public health partners – so that was our early Community of Practice. As we speak, we are going to present to the PHIN community. We're going to have, actually very shortly within the next few weeks, our first call. One of the challenges is, as I describe these tinker toys, [that] we got to make sure we don't have redundant tinker toys. Right now we want to make sure people are sort of aware of what everyone else is creating because it's a broad environment. So I hope that answers your question. It should be very interesting. Part of it is about our not directing what this community should do.

If it turns out they said, “Listen, you guys haven’t been working on this one area,” we’ll start to rotate. It’s really what the community needs.

ST. LOUIS: I think Henry and then Blake, while Henry is coming up.

BLAKE: Hi Tom, really nice job, fun to hear about. I’d be curious to hear how you expect to work with NEDSS. I’d like for you to elaborate a little bit on that. You had it on your list of collaborators.

SAVEL: It’s early with NEDSS. And what we’re in the process of looking at conceptually is in a sense about “gridifying” NEDSS. By basically by taking grid nodes to each on of the sites, that there’s a way, essentially to do a distributed query on NEDSS. We’ve been talking to them and working on some potential prototypes. So the answer is “yes,” and they’re very interested as well. Unfortunately, we had a NEDSS meeting that we had to cancel because I wasn’t feeling well, but we are back on track on that. I’m looking forward to really talking to them and finding out how it can really start changing the way they do business.

ST. LOUIS: On the phone or Envision. Otherwise Henry, please.

ROLKA: My question was quite similar to the last two in terms of a pilot or a program that sort of ready to be used with the grid. It sounds like you haven’t quite worked that out.

SAVEL: Let me cut you off on that. I would say “yes” in the sense that right now if you said, “We’ve got a data set,” we could work with your data set and say “Okay. Is there a subset of that data or a summary concept of that data that you are willing”…..oh, do you want me to talk about that. Okay, so BioSense. So we’re really looking at the transitioning of BioSense to a distributed grid model. And looking at finding it much more palatable for all the data – instead of being centralized – is that the data can remain where it’s located and these aggregate, essentially summery level data queries essentially are sent out to BioSense nodes. So yes, I didn’t

actually mention that. I'm very excited about that. The changes it can have in terms of the impact both emotionally. It's more palatable if all the data doesn't have to be centralized. That you can at any time do the query of how many cases of X are occurring in a certain location at a certain time and not have to have all the data centralized. That's currently in the BioSense strategic plan and it's in process.

ST. LOUIS: Other questions? Roy.

ROY ING: I'm Roy Ing, Informatics Consultants on Contract at the National Center for Environmental Health. One of the issues with all of this is that grid is like a road. We really have a lot of data at CDC, and as data providers, we need to have the most standardized way of making data available within ourselves to share info with the outside. I think we need to have something more like web services and so forth and make the data available electronically in some uniformed way, because right now, you can't even get, for example, peanut butter and *Salmonella* [data]. You can't get it in the electronic form although there posted as pictures of Epi curves and pictures of maps, but the data itself is not available. I think that is something we should all do as a first step regardless of whether you use grid or no grid.

SAVEL: Thank you very much for that comment. That's a very good point. For those out there in various centers, if they say, "What can we start to do?" I would say – start to think about if you have a system where you have data collection and you have some processing of that and you've got some sort of public health information system. Think about that if you had to transition to a much more of a service components where you were using web services for your system that would help get things ready, because we have tools that actually can start to transform standard web services into grid services. So, thank you for that.

ROY ING: But I think we need to have policies and procedures within CDC that will help the data providers prepare their data in a form and make it available. Because, right now, I don't know of any program that actually puts data out in a web service format. Some of them you can download: Excel comes separate,

which is great, but not all of them provide that. But we're talking about in this case we're talking about web services where another application, a computer program, can make a query and get that data in electronic form

ANONYMOUS: I think what you're referring to is the security policies as much as anything else and probably some other policies as well. One of the efforts that we're taking in the grid program is to certify and accredit the possibility of the Globus node so that people could publish web services via the grid. It is a long, laborious process, obviously, but it is one part of it we are working on.

ING: I'm not referring to the authentication or authorization. I'm talking strictly about having data available to the release. This is the first step that we can all do and that is something that will have to be done.

ST. LOUIS: Okay, great. Can we take a last call to see if there are questions on the phone? Or, elsewhere here? Please come on down.

CARL SOETEBIER: Hi, my name is Carl Soetebier. I'm with the Georgia Division of Public Health. I have a really simple, practical situation. Even though they're only a few blocks away from me, I cannot get data from my poison center. I noticed you have a service from the American Association of Poison Control Centers (AAPCC) in testing. So, I was wondering if you could just briefly describe to me what my experience might be in trying to use that service. And what are the steps? What does the participant have to go through to consume something such as that?

SAVEL: In terms of the process, we've been working with them I think in terms of essentially a user. So assuming you have rights to see the subset of whatever data from poison center, the goal would be that there would either be a login, user name, [or] password into a website which would then give you the access to look at that data or potentially if you had an application you already had built. You would do standard socalls but you would need a digital certificate to make sure there's that secure layer on the grid. So essentially, you would

– how should I put this – I think it was going to be system integration so it would probably be more a node. Go ahead Brian.

BRIAN: There are basically two approaches. As Tom said, AAPCC actually publishes a service. So you could contact them and try to get access to the service directly. Additionally, the application Tom demoed, the Quicksilver application, is for internal CDC usage initially, but it basically prints credentials in front of that service. Not everyone has a deal directly. And they can do national, state and zip code level data. Those two approaches are first up.

SOETEBIER: So then the enrollment process though is specific to the data provider…or, is there a global authority management? I guess that's where I'm at.

SAVEL: Currently, at this early phase, you would work probably directly with that provider. But in time – and this gets into a whole thing we can talk about – a lot of that having that overall security management framework and we are in the process of leveraging work that caBIG has done in managing countless numbers of nodes and certificates. By the way, it's called Guards and we're leveraging that for this Public Health Grid. Thank you for that question.

ST. LOUIS: Are there any further questions? If not, let me thank Tom once again and ask Ruth if she will come up to close down the seminar and give an update

JAJOSKY: I want to thank our speaker and our moderator for an excellent session. And I want to thank everyone who participated today, for taking time to contribute to this session. I hope it was of benefit to you and will spark further collaboration and work in this area. Our next seminar series will be Thursday, April 23rd from 12:30- 2:00 pm (Rescheduled for June 23rd from 1:00 – 2:30 pm EST). David Walker, Henry Rolka, and Janet Heitgerd will be giving a talk about BioPhusion. As always we invite your suggestions for future topics that

would be helpful in advancing integrated surveillance. And you may do this by emailing us at integratedsurveillance@cdc.gov. Thank you very much for participating.