Citizen Science Toolkit Conference

June 20 - 23, 2007

information commons: a catalyst for scientific and social innovation

Josh Knauer
Director of Advanced Development
Information Commons
MAYA Design

CORNELL LAB *of* ORNITHOLOGY

CORNELL LAB OF ORNITHOLOGY

607.254.BIRD telephone
www.birds.cornell.edu

159 Sapsucker Woods Road
Ithaca, New York 14850

This presentation took place at the Citizen Science Toolkit Conference at the Cornell Lab of Ornithology in Ithaca, New York on June 20-23, 2007.

Note that this document did not originate as a formal paper. Rather, it combines an oral presentation with accompanying PowerPoint slides and reflects the more informal, idiosyncratic nature of a delivery prepared specifically for this live event.

Documentation of the conference is meant to serve as a resource for those who attended and for others in the field. It does not necessarily reflect the views of the Cornell Lab of Ornithology or individual symposium participants.

The following is one of three focus point presentations delivered as part of the session titled "Technology and Cyberinfrastructure" on day two of the Citizen Science Toolkit Conference

For complete documentation of conference proceedings and to learn more about citizen science and the Citizen Science Toolkit, or to join the ongoing citizen science community, go to:

http://www.citizenscience.org

This documentation is supported by the **National Science Foundation** under Grant ESI-0610363.

Any opinions, findings, and conclusions or recommendations expressed in this documentation are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Information Commons: A Catalyst for Scientific and Social Innovation

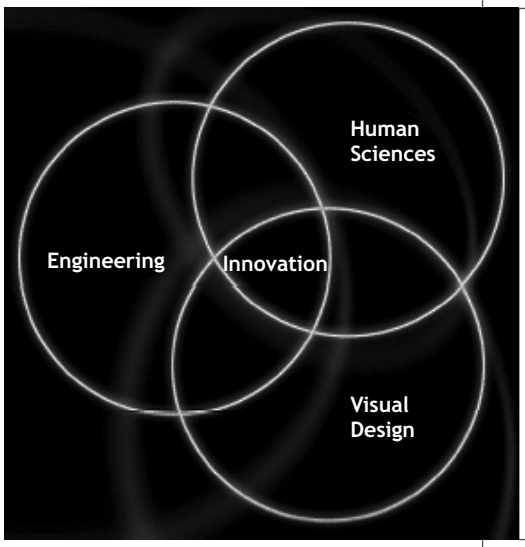
Josh Knauer,
Director of Advanced Development,
Information Commons,
MAYA Design

http://www.maya.com/infocommons

About MAYA Design

I'm going to begin with some quick background because I don't come from the citizen science community. I work for a technology research lab in Pittsburgh called MAYA Design. We take a multi-disciplinary approach to helping individuals and organizations understand how people interact with technology and the information that they get via technology.

A lot of our clients on the commercial side of our operation are Fortune 100 companies such as Panasonic and Whirlpool. I specifically work on the research side. We come from an academic background. Three professors left Carnegie Mellon University and started MAYA in 1989. We have a core focus when we look at technology, and think in terms of a concept that we call "information liquidity."



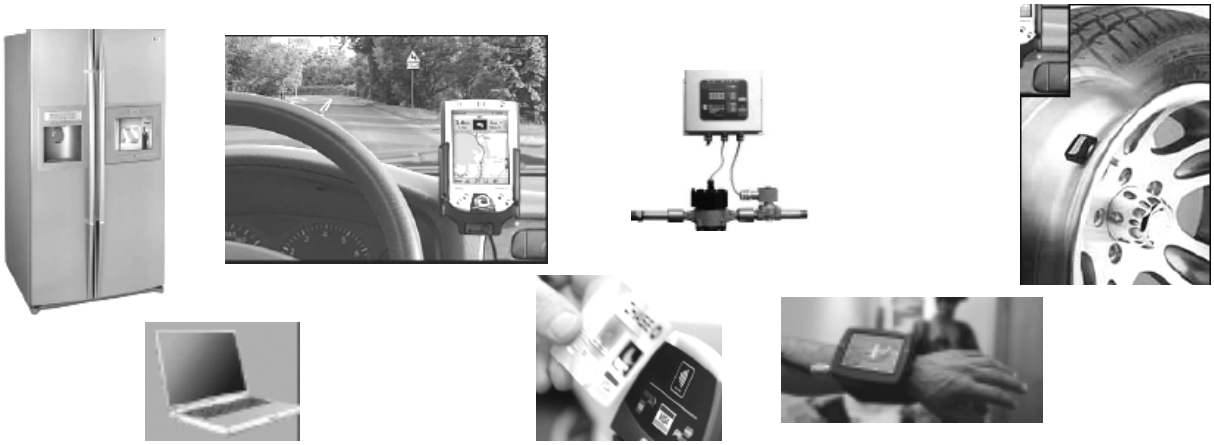
Background on Maya Design

- Technology research lab created by three Carnegie Mellon professors in 1989
- Multi-disciplinary approach
- $50 million in federal research to create technical architecture for "information liquidity"

The Big Problem: Information Liquidity

The concept of information liquidity came from thinking done in the late '80s by our founders. They thought about a crazy sci-fi future when there may actually be trillions of devices in the world that need to be interconnected. How do we have a single piece of data be available to all of those trillions of devices?

Trillions of Information Devices



Obviously, these are all devices that exist right now. We have computers embedded in almost everything: in our refrigerators, in our cars, on wrist computers. They're used in the military, but coming to a child near you soon. We worked with one of our clients on a system to embed a bit of computing and storage into every building system they sell, so there is the concept of pervasive networking throughout buildings in ways that we haven't even thought of how to take advantage of. And then, of course, there is the computer. In the work that many of you do, obviously the computer is a big focus point in terms of how people enter data into a computing device that we all link up to through the Web.

> "So the question isn't how you build a better Web forum for people to enter data, but how can you actually push your organization, your mission, out to devices that people are using and are going to be using in the next five to ten years?"

What I would like to pose to you is that in the future, computers as we know them are going to change, just as in the early part of the 1900s the concept of a motor or an engine changed. It was a very specialized concept that was in people's homes. You had big conveyer belts that ran off the one motor that everyone had in the home and it ran your lights, your fans, and so on. Now they're completely ubiquitous. Computing is going to be as well.

So the question isn't how you build a better Web forum for people to enter data, but how can you actually push your organization, your mission, out to devices that people are using and are going to be using in the next five to ten years?

As a result, the big problem that we've been focusing on is informa-

tion liquidity. Do existing information systems support trillions of disparate devices that are all in different formats and standards and all the rest? The question is, how do you get the data to where it is needed?

We also have a mission focus with our research, which leads to the question, how do you make sure that public data is always available at all times? If the government server goes down or if, god forbid, the EPA decides to change some of its data midyear as they've been known to do, how do you know what they've done? How do you get transparency of that information? We're talking about unambiguously public data.

Before you can think about actually doing all of this, we like to look back at history. We are big, big fans of history and looking at how technologies and other types of institution patterns have happened in the past.

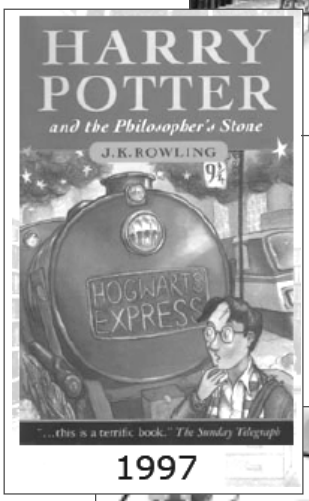
How do you get the **data** to where it is **needed?**

How do you make sure that **public data** is always **available**?

Are there any good historical models for doing this?

Public Libraries: The Original Information Commons

It turns out that public libraries are one of the best models for efficient data distribution or information distribution in a physical form. Let's explore this a little bit, and let's look at a very popular book from the Harry Potter series, *Harry Potter and the Philosopher's Stone*, which was released in 1997.



When any book is released into publication, libraries start to massively replicate copies of books. What happens is that they maintain the intellectual property of the publisher and the author, they maintain all of the information that you want to know about who published it and when and who the author is, and you get a complete copy of this intellectual property scattered across millions of libraries across the planet. At least I think there are millions, though maybe I should check that number. Let's just say "lots."

An interesting question arises when you go to a library that doesn't have the book that you want. This is where a very specialized function starts to kick in that the public library system in the United States and also around the world is very good at. In the United States you have

The Inter-Library Loan System saves the day!

something called the Inter-Library Loan System, so when you go to a specific venue and it doesn't have the information you were asking for or the book that you want, you are guaranteed that within a fixed period of time—a couple of days, maybe a week—that book will be sent to that venue by one of the other libraries out there. The intellectual property basically gets passed around and flows to where it is needed. I don't know if you realize this, but when an Inter-Library Loan happens, if it's not a rare book, frequently the library will then order a copy of the book that was just asked for. So it's a sort of demand model that happens in public libraries in the public sector that is very interesting.

This guarantees that there is always a copy of the intellectual property wherever it is needed. I was looking into the history of things like book burning, when you want to destroy things. I mentioned the EPA earlier. The EPA changes data—there are documents with proof of this—because of pressure from corporations that don't believe that the problems reported exist. Sometimes it's a legitimate problem, sometimes it's not, but the issue is that the record is destroyed, and if the EPA pulls it down off of its Web server, not many people are replicating that data sufficiently right now to have a true copy of what it was at a fixed point in time.

This guarantees there's always a copy…

Harry Potter book-burning, 2001, New Mexico, USA

…even when extremists try to destroy it.

In terms of book burnings, this incident in New Mexico was religious fanatics who have a problem with Harry Potter for whatever reason. This took place in 2001, so this isn't something that only happened in the past. The attack on intellectual property in the public domain side of things is very real, it happens all the time, but through replication the library system is able to basically withstand that. You can't burn all of the copies of Harry Potter. You can't destroy the concept of this book or any other that is stored within the system.

Applying this Model to the Digital Age

The Goal

The question is, how can this model be applied to the digital age? Right now, all of you who have Web sites and are storing your data in centralized systems and enterprise systems and all the rest, even though you are backing it up, are basically building massive silos of data. You are basically taking all of the important knowledge and putting it into one library without replicating it anywhere else. What is happening is that if you have a crash, if lightning strikes, if somebody malicious gets in, your data is lost, it's gone. That is a very big problem in the digital age and it is something that I worry about a

lot in terms of scientific knowledge and the dissemination of it in this world that we live in.

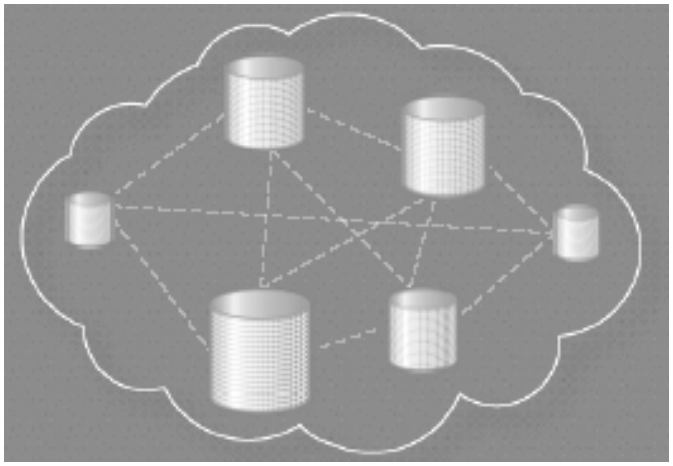
Our goal as a research lab is to unite all of society's public data and information into one open (and that is important), massively distributed database to ensure its availability to all. What we are proposing is really an information architecture. We call it a database because that is a language that you all understand, but what we're trying to promote here is a concept of how data can be shared and distributed and replicated across many, many trillions of machines and devices, similar to the way that the Web is an architecture for how information can be passed back and forth.

Our Goal:

Unite society's public information into one open, massively distributed database to ensure its availability to all.

The Need for an Information Commons

Our model for this literally mimics the public library system almost to the T. We have many, many venues—it could be a cell phone, it could be a laptop, it could be a drone flying around in the air, it could be your refrigerator or the light bulb. Basically anywhere that there is storage, a bit of computing and networking can be the equivalent of a public library and serve that same function in society, to help replicate and spread information around as much as possible.



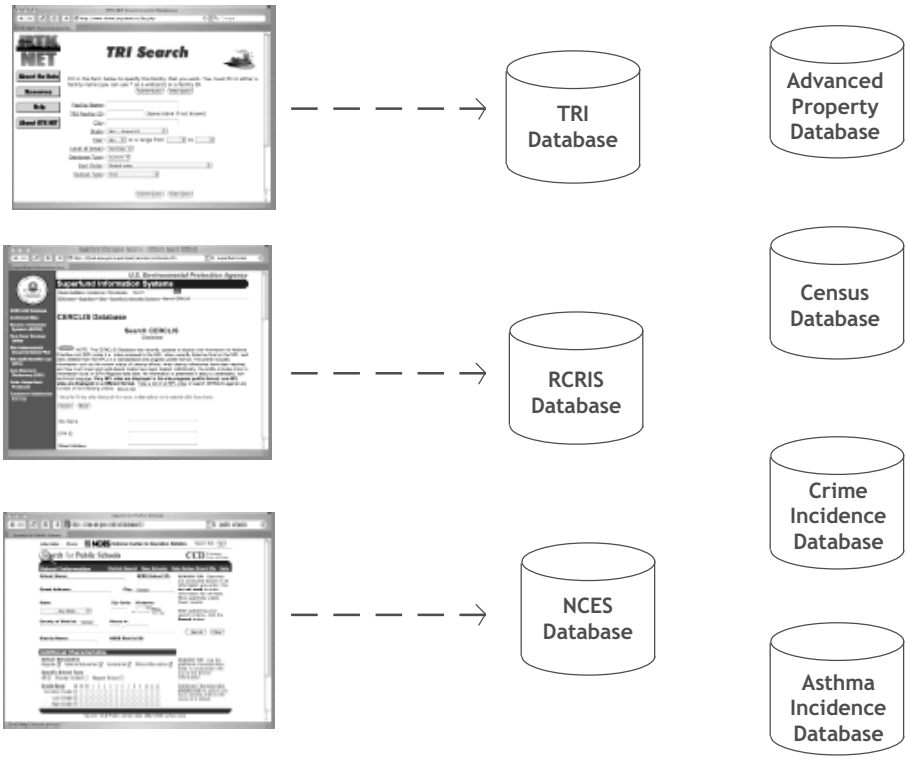
Information Commons

There are lots of other things that start to happen when you have one big database that everybody actually starts sharing. At this point, many of you may be scratching your heads, confused. Or you may be thinking, we already have the Web, we already have the Internet, what's the problem? Why are you trying to solve something that isn't really a problem right now?

To explain, I'll offer a very quick review of the current state of the art in terms of how we retrieve information. We do a lot of work with communities and the example I'm about to give you is one in which we started looking at the proximity of schools to toxic facilities in communities. Fourteen states have laws saying you can't build a new public school within half-a-mile of the toxic site in the community.

Well, if you're a parent and you want to know where the toxic sites in your community are, you've got some problems. If you're a data expert, a GIS weenie, you can try to do it using a limited number of datasets. You can go through the Web and download the Toxic Release Inventory, you can go and download the RCRIS database which gives you all the mom-and-pop storage of toxics in the community, like drycleaners and gas stations and things like that. Then you have to go to the National Center for Educational Statistics, which is a federal center where all of the data for No Child Left Behind gets sent.

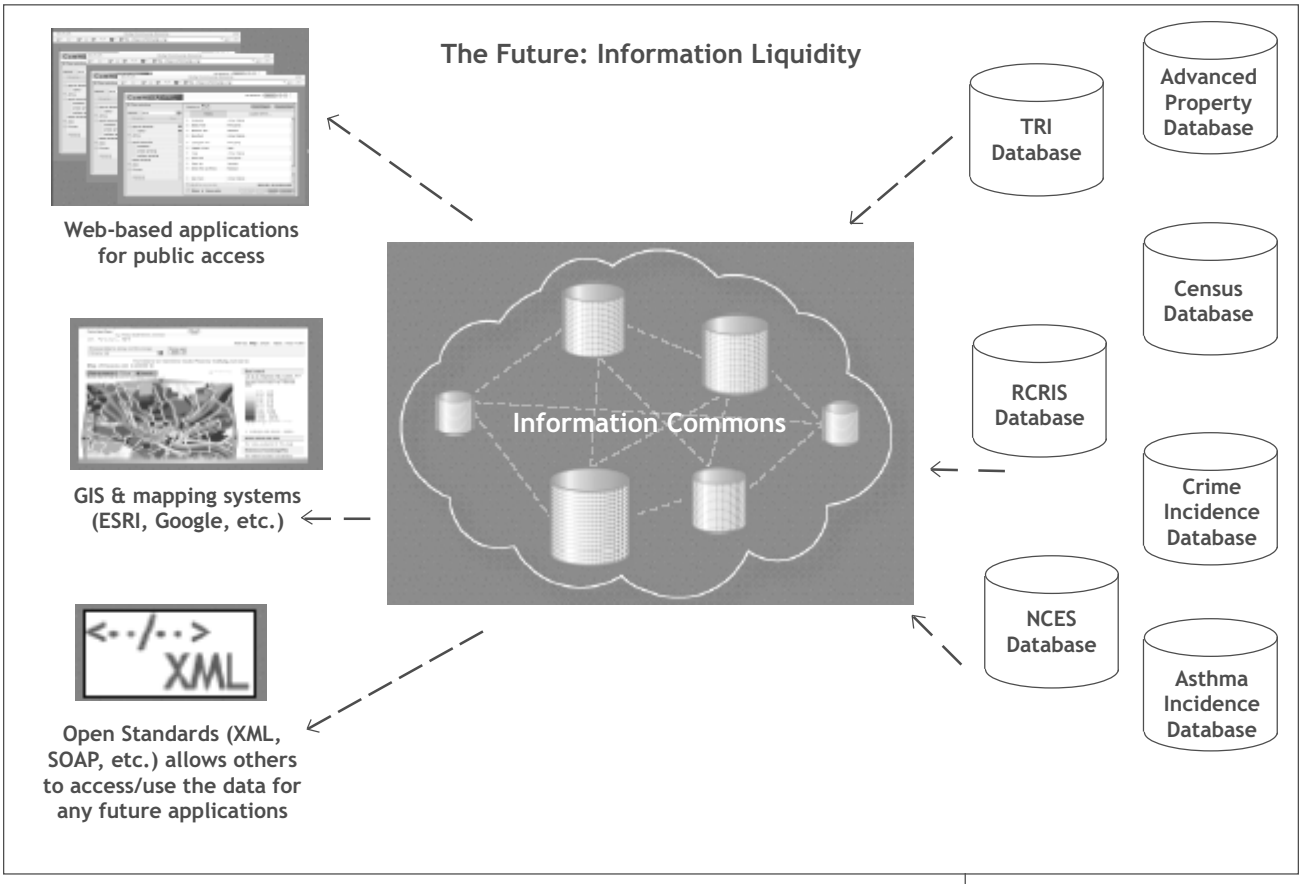
The Current State of the Art



The problem is when you start collecting this data and you say, "Wow, we're starting to locate schools right next to toxic release facilities," and historically we have, "what are the other impacts of these toxic facilities in the community?"

Then you start looking at property and the census and demographic type information, such as whether crime increases and all the rest, and you have a very big problem, even if you are a data expert. You end up building multimillion dollar Oracle-data-warehouse-type things that are not scalable across many, many different domains.

The issue that we thought about in all of this is, is there a different way? I have a limited amount of time for this presentation and I'm skipping over a lot of the deep technology architecture, so you'll just have to believe me that this works. What we've done is to take all of this data. We have thousands of data sets that we have collected from federal, state and municipal government entities as well as from organizations and individuals. We are starting to gather and transform that data into what we call a universal format. Let's not go there in detail, but it is a very simple data transformation that can take place from any type of data source, any data format, and transform it in a way that allows it to be replicated within this information commons.



Then what happens is that when you build applications, rather than going to the source dataset and trying to figure out how to download and fix the data because the latitude and longitudes are wrong, or whatever it is, what we do is allow many standards. The arrows on the left above represent user interfaces that are built through standards that exist today—Web standards, GIS standards, ESRI, and lots of other types.

There are also ones that we haven't thought about yet or that are starting to emerge. For anyone who believes that XML, SOAP, and existing methodologies for how we transfer data are going to exist five to ten years from now, and is thinking that we're far too evolved in this technology information infrastructure, consider that KML emerged onto the market and exploded out onto the market in less than two years. What if you asked people in the GIS community ten years ago what standard people would be using today? Today the USGS, the EPA and lots of other federal agencies are now starting to publish their data into KML instead of ESRI format because it's just easier.

That's going to evolve over time, so how do you as organizations figure out what to do and how to keep up with this? This is where information liquidity saves the day, as we have data transformations now that translate any data in the Information Commons into many, many different flavors of portable data formats like XML. The neat thing is that as we find out about new ones, a new standards release or whatever, it usually takes a coder a day or two to

grab that and transform the data into that format.

What is also neat is that because this is in a commons these arrows actually reverse, so basically you can start having data flow back in from sources that it's going out to. It isn't just a one-way flow of information. For example, if somebody builds a piece of software that allows you to transform data into some new standard that hasn't been developed yet, that software itself can actually be put into the commons and redistributed out and used by others. So not only can data be reused, but software and applications can be as well.

The Benefits

Through research partnerships that we have started to establish with other organizations, we've been finding that there are benefits to organizations in terms of using this type of architecture, and in terms of thinking about their data as a fluid resource and the organization as a fluid resource that can flow out onto many, many different computing devices.

Probably the first and foremost among them is the reuse of data. We did a project with the Heinz Endowments in Pittsburgh looking at environmental toxins and how people come into contact with them. One of the big problems that they came to us with was that in 2004, many different organizations came to them for funding to go out and hire GIS consultants to download the Toxic Release Inventory so that they could do an analysis for toxins in watersheds.

The interesting thing is that of that funding, which happened many different times, we calculated that sixty percent of the work was duplicated. By having a common shared resource of unambiguously public data like the Toxic Release Inventory, where somebody else has already done the work, you should be able to reduce duplication.

Benefits of Using the Architecture

- Reuse of data and code
  - across many applications, projects, organizations
- Opportunities for data fusion
- Collaboration among domains/organizations
- Replication…it's always available
- Incremental growth and development

By the way [*referring back to the diagram, top of page 7*] you can do things like attach what we normally think of as metadata to the data on the way in so you know who put it in. They digitally sign it and when they put the data in, they include what the source of it was and so forth, and if you choose to trust those sources of the data, you can then basically filter. You could see that three different organizations have imported a Toxic Release Inventory for whatever reason, and you could choose which one or which groups of them you want to trust and filter into the system that you built. If you trust other institutions with some of that data and you reuse that data, you can have an exponential savings of time,

cost, and effort in integrating your effort with other domains.

We are also finding it is not just about the data. It's the reuse of the data and the reuse of the code, and we are seeing this happen across applications, across projects, and across organizations now over time. You should all be very happy that two years ago the University of Pittsburgh took the entire bicennial census for 2000, 1990, and 1980, and imported all of it into the commons down to the census block group level for the entire United States. We have many different projects right now that are incorporating that data into their projects at almost no incremental cost because it was just there in a format that they knew. They could put it on their Google Earth site or their Web site or their custom application, or they could import it into ESRI if they wanted to. So data reuse is really big.

"We are also finding it is not just about the data. It's the reuse of the data and the reuse of the code, and we are seeing this happen across applications, across projects, and across organizations now over time."

Then there is another benefit, and this is a controversial one for a lot of people. There are opportunities for data fusion. For example the Cornell Lab of Ornithology may collect the bird count data that they have and want to organize it into counties and identify how many birds per county were found. We have lots of data about counties that we have collected from other places. We have census data about counties, we know all the schools in the counties, and we know all the nonprofit or tax-exempt organizations by county. The whole point is that we could actually cross-correlate populations of the American Robin to religious organizations in a given community. Hopefully every scientist in the room is protesting, "Oh my god, you can't do that, that's not why we collected the data."

The interesting point is that you could try to see if there are correlations between that data in places where you never could before. This actually led to something that was very interesting. We did work with the highly endangered Florida panther and collar readings from the panthers, and made wonderful visualization tools with the very precise locations of them. This is where you start stumbling into the social mistakes that start happening around data: Just because we're geeks and we can, we think we should, right? That is why I now frequently talk about "unambiguously public data" as opposed to highly confidential information. If you're talking about the last eight Florida panthers on the planet, you don't want to be providing the exact GPS collar readings of where they might be found.

What we did at the time was to correlate that data to breast cancer mortality data. This was me just showing off random data sets seen together. We were doing this at a center for oncology and one of the doctors said, "That's preposterous! You should not be doing that!" It wasn't because of the endangered species issue, it was because it was bad use of data and cross-correlation.

The doctor standing right next to him said, "Well, do we know that

there is no cross-correlation?" Obviously you can do cross-correlations all over the place, but you have to be very careful about how you do it and we have a lot of experience with that.

This allows for collaboration across domains and organizations, as I've said, and I've already pointed out the replication issues. The most important issue here is the incremental growth of your system. This is the thing at which we all fail. We set our rockets up for the moon shot and we get millions of dollars of funding from NSF and the foundation community and all the rest, and something happens and we forgot the attribute that we needed to add into our application and usually we have to redo the entire thing. It's a very expensive process. Moon shots are a bad idea because frequently the geeks get it wrong (sorry, but we do), and also frequently the people collecting the data get it wrong as well. You want to be able to adapt over time the types of information you're collecting and the other types of correlation you want to be able to make with the data. Our system, in the way that it deals with ontologies and things like that, allows for that to happen almost to a fault. A lot of people find it hard to get their heads around the fact that they can do that.

Lessons Learned

I try to stay away from using lots of language that geeks tend to use, but one of the lessons we've learned along the way in terms of how to get information to start flowing and become liquid is to separate identity of the data object, the thing you're trying to describe in your data structure, from semantic meaning. For example, the big thing in science right now is the Semantic Web, and in the Semantic Web world, they tend to identify individual data objects by the location of the server it's on and the location on the server that it's on. And trust me, I've had very animated discussions with people in the Semantic Web world about this and they claim it's not true, but in practice, that's how it happens. The problem with that is, what if that server goes down? In a distributed world, going back to the public library system, you don't care where the library comes from, you just want to make sure you get your data.

We very much believe that semantically meaningful identifiers, like names for example, are very bad. For example, if you go to Wikipedia, lots of the concept names and place names are in English. Well, lots of the world doesn't speak English or express themselves in that way and there is a very hard effort to try and cross-correlate between multiple translations of a word and the concept that it refers to in Wikipedia. In fact, the only solution they've come up with is to create separate

Lessons Learned Along the Way to Liquidity

- Separate identity from semantics
- Peer to peer distribution
  - open architecture
  - one database across every device
- Metadata IS data
- One data object can belong to many ontologies
- Allow arbitrary incrementalism

duplicate listings for every translation of the same concept.

Peer-to-peer distribution is very important. This is an open architecture. MAYA built this, we came up with the architecture for this, we published the architecture so that other people can build their own versions of the database application that runs it, similar to the way there are many different types of Web servers (Apache, Microsoft IIS, etc.). It is already at the point where there is no centralized control of this commons. There are enough organizations that we partner with right now, and enough data flowing around and replication of that, that I actually believe that it couldn't be shut down. It's an interesting concept—it's completely out of our control at this point, which is great.

The real key concept of it though is that there is one database in the world. And this is the future, whether it is this architecture or another, and it has to be distributed. It can't just be Google or Yahoo or any one company that does this for us, it has to be openly distributed everywhere.

Another thing that we've learned is that metadata itself is data. We don't separate metadata from the data, it is part of it. You can take any individual piece of data out of this information commons cloud that I was showing you, and you can know everything about it. That is very important because you need to be able to mix and match data across multiple data sets and know the impact of that.

The next lesson is one that the ontology people like: One data object can belong to many ontologies. You don't have to design the uber, be-all-end-all ontology. You can build incrementally and craft the data incrementally across ontologies. That means classification systems of geography, for example, or species, or whatever. When you are talking about the American Robin for example, if some child or group of children decide to call it the "orange rusted bird" and that is the way that they classify it, that's okay. Somebody can make a mapping of that, publish that mapping back into the information commons, and you can start ascribing lots of other assertions that people make regarding orange rusted birds or American Robins and have multiple ontologies converge.

Further Information

We have many papers written about this. I encourage you to read some of the papers that we've written and others have written about us and this process. Almost all of them are at the Web address at right, and a lot of the work that we've done has been peer reviewed, so feel free to dig in and ask a lot of questions if you want to.

We're Looking for Partners!

For more information:

http://www.maya.com/infocommons

Josh Knauer
knauer@maya.com
412-488-2900