The National Infrastructure for Community Statistics: Liberating Public GIS and Statistical Data

Pari Sabety
The Brookings Institution

Chris Sweeney
3 Rivers Connect

Dominic Widdows, Joshua Knauer, Maryl Curran Widdows, Peter Lucas
MAYA Design, INC

August 29, 2005

Appeared in the 4th Annual Public Participation GIS (PPGIS) Conference, Cleveland, Ohio, July 31 – August 2, 2005.

Abstract

The needs of local communities for Comprehensive Information Systems (not just Geographic Information Systems) require new social, political and architectural models that enable collaboration and the seamless integration of all types of data into an Information Commons, breaking down the barriers between expert and non-expert users, and between geo-data and non geo-data. This paper describes the Pittsburgh Community Information System (CIS), a pioneering collaboration involving 3 Rivers Connect, the Universal Database technology of MAYA Design, and the National Infrastructure for Community Statistics project sponsored by the Brookings Institution. The Pittsburgh CIS enables desktop access to a massive distributed database of real-time, up-to-date information from federal agencies, local agencies, and non-profits, that non-technical users can query and analyze with powerful, easy-to-use statistical tools and GIS mapping capabilities. The Pittsburgh CIS is part of a growing collection of integrated public datasets called the *Information Commons*, implemented using a Universal Database technology that enables the crucial goal of *information liquidity* to be realized.

Introduction

Spatial information is supremely important to the way humans understand the world around them. Graphic user interfaces enable us literally to *put* information *in* files and folders, and to organize our desktops and workspaces according to our personal needs and tastes, even if the information arranged in this fashion is not naturally spatial in nature, and the spatial organization is really a convenient metaphor. And yet, for data that *is* naturally spatial in nature — that is, geographic — the organization and manipulation of data is still largely the province of specialized professionals, and relies on costly GIS systems that are not available to the public at large.

To some extent, this is changing in some simple commercial arenas. for example, Google and Microsoft are both producing "earth" models, which are catering superbly for anyone

in the world who doesn't know where to buy a cappuccino, but these simple consumer products may never cater for the needs of true community information systems. Locating community services, the routing of buses, the fixing of school district boundaries and catchment areas, are examples of hundreds of interconnected problems that affect the whole community. However, those trying to solve these problems often have inadequate access to the information they need: though this information may exist in some form in some data venue, it is often constrained in many ways, and it is fair to say that, in traditional systems, no community decision-maker ever has access to all of the data they need in the form in which they need it.

To solve these problems, it is necessary to connect information, while allowing for many different ways of selecting and interacting with that information. It is time to move beyond data silos, relational tables, layers and architectural models that perpetuate the incompatibility of geo-data with the realm of socioeconomic statistics. The needs of local communities for Comprehensive Information Systems (not just Geographic Information Systems) require new social, political and architectural models that enable collaboration and the seamless integration of all types of data into an Information Commons.

This paper describes the collaboration of three groups in solving these problems: the National Infrastructure for Community Statistics, MAYA Design, and 3 Rivers Connect.

A project of the Brookings Institution's Urban Markets Initiative, NICS is an unprecedented public-private, federal-state-local collaboration facilitating an information marketplace to exchange data, tools, information products and technologies, liberating information to improve decision-making at all levels of government and stimulate economic development.

NICS is collaborating with a number of local data-providers and enterprises, such as 3 Rivers Connect (3rc), a non-profit dedicated to creating a Regional Information System for southwestern Pennsylvania. 3rc's vision is to create a regional information asset for southwestern Pennsylvania that will allow area organizations to dynamically create, combine and retrieve vital information, regardless of its original location. In this mission, 3rc has paved the way in negotiating the politics of information sharing, and developing best practices for creating effective collaborations between the non-profit, government and private sectors.

The collaboration between NICS and 3rc is facilitated by the technological contribution of MAYA Design, Inc., a Pittsburgh research and design company dedicated to taming the complexity that has exploded in the information age. MAYA specializes in creating information systems that are tailored to the needs and opportunities of the future, such as data liquidity, not the constraints of the past, such as fixed data locations and inflexible schemata.

This paper describes the contributions and collaborations of these three organizations, in the context of providing Community Information Systems in the city of Pittsburgh and the wider region of southwestern Pennsylvania.

NICS

The National Infrastructure for Community Statistics, or NICS, is a collaborative community dedicated to creating an environment for mutual information sharing to enable better

community analysis and decision making. The NICS concept and the NICS Community of Practice were created through the efforts of the Community Statistical Systems (CSS) Network, a loose community of public, nonprofit, and university actors interested in the spread of community statistical systems around the U.S immediately following a major conference organized by the CSS Network, held in Reno Nevada in 2003. Some of the major conclusions and the guiding principles behind NICS are discussed by Reamer and Sabety [2003].

Limitations of Current Community Information Systems[1]

Local community statistical systems are proliferating rapidly around the country, primarily in metropolitan areas, and are complemented by federal and state counterparts (e.g., FedStats, state business and economic research centers). These systems represent a revolution in the public availability, display, and use of data for local areas. This revolution has enhanced the ability of a wide array of data users, from neighborhood groups to Mayor's offices to national policymakers, to understand local socioeconomic and geophysical conditions, identify trends, and measure program and investment impacts. Overall, users have been able to paint a more multi-dimensional, integrative picture of local areas than was possible in the past. In particular, community statistical systems have enabled community-focused organizations to undertake initiatives to build assets at the community level more effectively and implement multi-sector strategies to alleviate poverty.

Community statistical systems also have demonstrated an important contribution in urban markets because they allow private sector decision-makers to better analyze markets, identify opportunities, and make effective investment decisions. Consequently, the growth and development of community statistical systems offers significant positive potential for facilitating commercial business activity and employment in urban neighborhoods across the United States.

While the emergence of community statistical systems reflects major advances in information technology and innovative approaches to government function, this tool is coming up against significant limitations:

- The nationwide network of community statistical systems is highly fragmented. Each system is designed and operated independent of others, resulting in duplication of approaches and scarce resources wasted on redundant routine data services and operations.
- Community statistical systems are constrained by the "data center" model, hosting datasets from multiple sources on a central server.
  - This model is too easily overwhelmed as the number and spatial resolution of relevant datasets greatly increases.
  - Copies of records on the data center server cannot be easily updated as changes occur.

[1] Dr. Andrew Reaner, NCS facilitator, contributed significantly to this section of the paper.

- As many individual data sets are ill-matched with one another, reconciling and re-interpreting these records at each data center is difficult, costly, and time-consuming. Moreover, users lack the capacity to customize data sets for their particular interests and have difficulty integrating "official" datasets with locally generated knowledge about community assets, habits, interests, and opportunities.
- The metadata inconsistencies stemming from fragmentation across the network of community statistical systems make integration and comparison of data from multiple sources difficult.
  - A system in one metropolitan area cannot carry out inter-metro comparisons using local data produced in other areas.
  - Data from different levels of government — state, local, federal — cannot be integrated easily.

In summary, the high resource requirements of the existing CSS model significantly inhibit the ability of local organizations to build community statistical systems.

However, while the above issues are substantial, they are not insurmountable. In fact, technical and institutional approaches for addressing them — such as adopting metadata standards, building distributed systems (federated data repositories), writing intelligent middleware, and creating partnerships among data providers and users — have been demonstrated. While these efforts are in early stages of development, they strongly suggest that with a unified technical approach and multi-institutional collaboration and investment, current limitations can be overcome. More specifically, these advances allow us to conceive and initiate development of a National Infrastructure for Community Statistics that would facilitate a significant increase in the number and capacity of community statistical systems across the U.S.

The Importance of NICS to Community Development

The implementation of NICS will greatly enhance the capacity of community-focused institutions — local governments, community-based organizations, community foundations, and universities — to ascertain metropolitan, city, and neighborhood conditions and dynamics in detail and to evaluate the impact of specific interventions. The huge value of accurate information in evaluating and making decisions is described by Sabety and Carlson [2004].

NICS represents the next wave in the evolution of public electronic access to community statistics. Recent advances in information technology have enabled the development of community statistical systems, local electronic intermediaries that provide public access to data on multiple topics (e.g., demographics, environment, employment, health, education) from multiple sources (including federal, state, local, nonprofit, commercial). Typically, community statistical systems combine access to datasets with functionalities such as customized table- and map-building and statistical analysis.

NICS is a web-based data services marketplace intended to provide participants with access to:

- thousands of community-level socioeconomic datasets on multiple topics (e.g., population, education, jobs and businesses, environment, health, public safety) from multiple sources (i.e., government, nonprofit, commercial),
- software tools to view, integrate, transform, and analyze these datasets, and
- resources for constructing web-based community statistical systems (data intermediaries) that bring together datasets and tools in one place.

NICS will be based on a service-oriented architecture, a "loosely-coupled" web services platform. "Loosely coupled" means that any web service — datasets, tools, or resources — can be made available on NICS in a form close to their current format. Following certain protocols will be required for participation (e.g., fully describing metadata). Some basic metadata standards will be encouraged to facilitate compatibility, and tools will be sought that enable the integration of heterogeneous datasets. Guidelines will be provided to describe the characteristics of "NICS ready" data services.

One analogy to NICS is a stock exchange. Companies listing their stocks need to meet certain criteria before doing so, but the stock exchange itself is somewhat agnostic about the types of companies listed. Within defined boundaries, market demand guides the nature of the content.

It is expected that NICS participants will include data intermediaries (e.g., community-based organizations, national trade associations, state data centers, federal agencies, commercial data vendors) and federal and state statistical and program agencies with the mandate to gather and analyze administrative data from local governments across their respective jurisdictions. The community data flowing through NICS will power thousands of data applications and web tools for end users.

The development of NICS will greatly expand and accelerate the proliferation of comprehensive, useful local community statistical systems across the U.S. Specifically, NICS will lead to:

- greater cost-effectiveness in the construction of local and state data intermediary web sites by reducing the need for repeated investments in labor-intensive operations such as attaching metadata, data cleaning, and homogenization;
- greatly increased access to useful local, state, and federal sources of community-specific data;
- innovative software tools for facilitating data analysis, performance evaluation and program assessment;
- easy access to benchmarking data from comparative communities and programs across the nation;
- increased data user awareness of potential data sources and uses;
- increased communication and exchange of best practices among community-focused data intermediaries and data users across the U.S.; and

- a broader, more vocal, more organized constituency for adequate funding for federal, state, and local statistical efforts.

The resulting growth and development of local community statistical systems will help communities and governments greatly improve the understanding of local social, economic, and geophysical conditions and the quality of discussion, analysis, choice, and evaluation regarding the investment of scarce public and private resources. For example, NICS will allow communities to:

- respond to changing demographics and socioeconomic conditions;
- use data, rather than anecdote, to better promote positive dialogue between local government and residents;
- invest scarce resources more efficiently; and
- develop more effective tools to adapt to economic forces of change such as trade, investment flows, and globalization.

Information Liquidity: A New Concept to Power Loosely-Coupled Systems

Those pioneers who are building infrastructures like NICS through a community of practice, or collaborating with Pittsburgh's regional information system, are struggling to tame the complexity of actually building a "loosely coupled" system in a way that is extensible and replicable across domains and jurisdictions, and can withstand the inevitable migration of technological standards over time. Addressing this challenge requires a new way of thinking about information — not as bits and bytes, but as liquid. In the past, investment in public information systems has tightly coupled data collection, data access, and user interfaces. The nature of political decision-making has often much preferred projects that promise to deliver standalone systems which can be identified with a particular administration, rather than systems that are flexible enough to be future-proof.

However, as the Internet has gone through several different stages of increasing maturity, it has become increasingly clear that the greatest long-term benefit of prudent investment in information systems is in making data permanently available, even though interface technologies and funding situations change every year. Data architectures are needed that encourage several projects to reuse one another's data, to submit updates and change-requests to preserve and improve data-quality, and to find new uses and new value in previously collected data, even though these uses may be quite different from the intentions of the system for which the data was initially collected.

The goal of information liquidity is to enable data reuse and data aggregation by encouraging data to flow between different machines and different projects, so that all users of data can have the information they need at their fingertips, indexed and organized in the way their users need.

Some systems are clearly more liquid than others. For example, fixed-schema relational databases are extremely non-liquid, because the identity (primary key) of a piece of data is only relevant in the context of a single data table, and can only be accessed by a system that

can directly call this data table, and understands the fixed schema. Moreover, it is normally impossible to add new variables to the original schema without changing the entire data table.

A much more liquid system is XML, with its extensible schemata and the potential that many different systems may read the same XML file and recover different information according to their needs. However, current systems based upon XML still have drawbacks — for example, moving data to a different location changes its identity. If I obtain a copy of an XML file from a server with a fixed URL (Uniform Resource Locator), the file I receive is merely a copy, and if the file becomes unavailable at its original location, there is no reliable way for new users to search the web for copies that were made before the original server went off line. For the same reason, systems based purely on Web Services models often have to run the same remote queries every time they want to receive the same information, which imposes an extra burden on the providers of highly respected, frequently used data.

It is therefore very interesting to consider systems where the data is more liquid still than is usual with XML technologies: where the data can flow throughout the system while retaining its original identity. Imagine a global information system that is as flexible and extensible as XML, but also provides automatic mirroring of data everywhere where it is needed, and manages automatic updates so that if a publisher updates their information, these updates propagate automatically throughout a global information space. In such a world, data is more readily available, and organizations who invest in collecting and publishing high-quality data do not get penalized for this effort by facing spiraling bandwidth costs.

The Universal Database: An Architecture for the Information Commons

Such a futuristic information architecture, though in its infancy, has already been successfully deployed by the Pittsburgh Community Information Systems that are discussed in the rest of this paper. The Pittsburgh Neighborhood Information System is built upon the *Universal Database* architecture of MAYA Design, Inc. MAYA Design was founded in 1990 by three researchers from Carnegie Mellon University, who combined the disciplines of computer science, industrial design, and cognitive psychology to create a successful research laboratory and design consultancy.

Much of MAYA's research over the past 10 years has been the design and implementation of arbitrarily scalable information systems, culminating in the development of an architecture that is proposed as the *Universal Database* architecture of the future.

U-forms

The Universal Database is constructed from a single abstract datatype, called a *u-form*.U-forms are described in more detail by Lucas and Senn [2002], Lucas et al. [2005a]; a brief overview of the concept is given here.

U-forms are generic data-containers: that is, any kind of data can be kept in a u-form, and u-forms can be accessed from any venue in the Universal Database using an extremely

$\sim$ 013a64ab30588c11d6b09a4cf30b756185

`name`	Pittsburgh
`country`	US
`state`	Pennsylvania
`latitude`	40.44
`longitude`	-79.996

Table 1: Simple u-form representing the city of Pittsburgh, showing the UUID and a few attributes and values.

simple interface. A u-form is simply a bundle of name-value pairs associated with a universally-unique identifier (UUID). A basic example of u-form may be the representation of the city of Pittsburgh given in Table 1.

The idea of collecting sets of attributes and values into distinct bundles representing particular objects is due to Dertouzos [2001], who proposed this architecture for an "electronic form" or *e-form* that computers would use as a universal means of communication. MAYA's innovation on this structure is the addition of the UUID, which supports the need for persistent identity and reference of information objects. Adding this unique reference to an e-form gave the name *u-form*. The challenge of producing a UUID for each information object, and guaranteeing its uniqueness, is often regarded as an impossible burden by newcomers to the system. In fact, numerous successful methods exist for doing this, as described in detail by Lucas et al. [2005a]. For example, it is possible to take a combination of the unique hardware address of a computer and the time given by the system clock, to create a UUID that can be thought of as a unique set of "spacetime coordinate" in cyberspace. This technique has been used for some decades by Open Source programmers and by Microsoft to give unique identifiers to objects, and has recently been proposed by Leach et al. [2004] for incorporation into the URN system of internet namespaces, a development we would applaud. Far from being the weak point of the system, generating UUID's is very easy, because we are no more likely to run out of UUID's than we are to run out of numbers.

Shepherding and Replication

Since the identity of the u-form (its UUID) refers directly to the information object, not to its physical location, the u-form abstraction enables copies of u-forms to exist simultaneously in many different repository venues. The act of creating a new instance of a u-form in a different venue is called *replication*. Replication is a great bonus for data liquidity, because it enables frequently requested u-forms to exist in many locations near to where they are needed, avoiding delays, heavy network traffic, and unavailability of information when servers go down.

The task of replication is carried out by artificial agents called *shepherds*. The task of shepherds is to make sure that, when a user edits a u-form in one venue, those changes propagate to all other replicates of this u-form.

Metadata in the Universal Database

In contrast with relational tables, a u-form does not have to follow any predefined schema when it is created, just as an XML document does not have to have a predefined set of permissible tags. New attributes can be added at will (this practically defines what it means for a data-object to be extensible).

However, it is extremely important for attribute names to be chosen so that u-forms can be interpreted according to the intentions of their creator. For example, the Pittsburgh u-form in Table 1 uses the attributes `latitude` and `longitude`, recognized by MAYA's mapping applications as global coordinates. If the creator of the u-form uses `lat` and `long` instead, a rendering agent that is looking for `latitude` and `longitude` will not work.

Metadata information about the intended interpretation of their attributes is encoded by metadata-carrying u-forms called *roles*. A role is simply a u-form that asserts that, if a particular attribute is present, the creator of the u-form intends that the value of this attribute should be used in a particular way. To use a particular role, a user simply adds a relation to this role's u-form in its *roles* attribute. In this way, u-forms carry relations to their own metadata schema. The roles system is described in more detail by Lucas et al. [2005b].

For example, the real Pittsburgh u-form includes a relation to the *Geopolitical Entity Role* in its list of roles. This role asserts that the value of the `name` attribute should be a human-readable string, the value of the `country` attribute should be a relation to the u-form representing the country in which the city is located, and that the `latitude` and `longitude` attributes should be interpreted as curvilinear coordinates describing a location on the planet Earth. (By default, these coordinates are interpreted using the WGS-84 reference ellipsoid, though the *Role for Georeference using Global Coordinate System* enables other datums to be specified.) In this way, we see the beginnings of ways in which GIS information can be expressed in a general-purpose information system.

The Information Commons

While the Universal Database architecture has many commercial and government uses, its biggest potential (and our biggest goal) is to fuse the diverse array of publicly available data into a single and diverse *Information Commons*. The Information Commons will provide a public store of information, together with the universal references necessary for that public information to be understood and fully utilized.

One of the best ways to describe the power of the Information Commons is to give a list of some of the datasets that have already been collected. A varied sample includes:

- World Countries and Administrative Subdivisions
- Shapes of Countries
- World Coastlines, Detailed (GSHHS)
- World Populated Places
- US Census Data

- Transportation Data. Roads and Bus Routes for different regions
- Major Airports
- Military Bases
- Land parcel data for various counties
- Chemical Substances
- Toxic Release and EPA Data
- Human Services Data. Facilities, Providers and Services
- Literary Datasets
- Lexical Databases

The power of the Information Commons is not, so far, the creation of these datasets — the world at large is doing a marvelous job of creating enormous amounts of data for a variety of disparate purposes. The value of the Commons is that these datasets have all been collected into a single namespace. This enables different datasets that are normally encapsulated with incommensurable syntax and semantics to be combined to make new and powerful inferences.

The process of collecting information and merging it into a growing structure is called *data fusion*. Data fusion is a difficult and varied task, involving all sorts of indexing, merging, spatial and textual inference, and genuine semantic sophistication. While we have many mountains still to climb in this arena, there are major causes for optimism. The single most important factor working in our favor is the universality and persistence of all data that is expressed in u-forms. Because all of the data listed above now exists in u-forms, it is possible for any of these datasets to refer to any of the others, knowing that the reference will be persistently valid, and knowing that the data object in question will be universally available even if its original venue has long since ceased to exist.

The Information Commons and GIS

One of the most important barriers that the Information Commons breaks down is the barrier between "geographic" and "non-geographic" data. In most current systems, data is either non-geographic, and will not appear on a map, or is geographic, and requires a costly professional GIS system (and engineer) to manipulate. We believe that this distinction is artificial and unnecessary. Building upon earlier work described by Lucas [2003], we have built interactive platforms that enable ordinary users to manipulate data traditionally regarded as "geo" or "non-geo". For example, it is easy for a user to create a u-form to represent themselves, their name, date-of-birth, and place-of-birth, and, by linking to the appropriate populated place u-form, their "birth-event" can easily be placed on a map, and on a timeline.

A practical example of such an interface in action is depicted in Figure 1. This map displays a variety of information collected during a project focused on tracking the behavior of the endangered Florida panther, correlating the range of the panther with Federal and

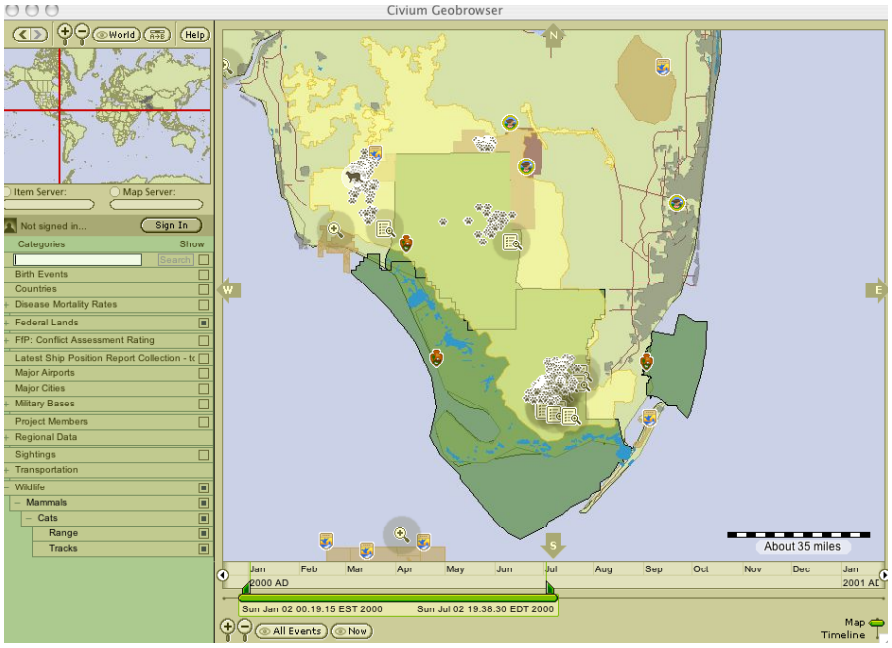

Figure 1: Visualization of sightings of the Florida Panther

State protected lands, urban sprawl and industrial effluent data. One of the most surprising things about this interface is that all of the datasets depicted are represented in u-forms — coastlines, roads, parks, industrial facilities, and the whereabouts of the panthers themselves. This, and many other examples, demonstrate the power of the Universal Database architecture in expressing data from a huge variety of sources, fusing this information into a coherent picture, and displaying this information to enable a human user to visualize the situation and make decisions.

Some of the scientific details of the Information Commons approach include the following:

- Point data may be expressed in latitude / longitude / elevation coordinates or with respect to Earth-Centered, Earth-Fixed Cartesian coordinates [El-Rabbany, 2002, p. 51].
- Shape data is all expressed in Cartesian coordinates. This enables shapes to be easily transformed between different coordinate frames, and enables one shape to easily include another by giving a UUID reference and a coordinate transformation from the parent shape to the child shape.
- This even enables a GIS system to incorporate shapes from a 3-d CAD-type model.
- Many real-world object share shape data. For example, the border of the USA on a map is also (in part) the borders of Canada and Mexico, and also (in part) the border

of the North American continent. The world-model fused from country and coastline data by MAYA Design encourages these shapes to be shared between objects, enabling greater resolution and data quality, while at the same time leading to improved performance through lower network traffic.

Publishing to the Information Commons

Far from presuming that all contributors will agree on what data should be regarded as definitive public information, the Information Commons provides a means for many different points of view and assertions to be expressed: but with respect to a common frame of reference. In broad strokes, the architecture for publication to the Information Commons is as follows:

- There is a core collection of *definitive* u-forms. These serve the purpose of defining an entity in the real world, just enough for users to agree that they are using a common frame of reference. For example, the u-form for Pittsburgh in Table 1 says nothing about whether Pittsburgh is a good or a bad place, or what its most attractive features are: it gives simple attributes like name, state, latitude, and longitude, which are enough for users to find the u-form for Pittsburgh on a map or using a search engine.
- Any user can comment on any definitive u-form, by creating a u-form that contains a relation to the definitive u-form.
- Many such u-forms are contributed by specific publishers. For example, the University of Pittsburgh Center for Social and Urban Research (UCSUR) is the publisher of US census data in the Information Commons.
- Users can select the publishers who they trust and whose information they want to see. Thus, if a user wants to see the census data about Pittsburgh *or any other city*, the user subscribes to the *census* channel of the publisher UCSUR.
- Interfaces dynamically adapt to this content selection. Thus, while there is a single definitive u-form for Pittsburgh, Pennsylvania, there is a potentially infinite variety of interfaces whereby different users can see different information about Pittsburgh provided by different publishers.

MAYA believes that this model is particularly appropriate to the needs of NICS and many of its member organizations, since it enables service-based, loosely-coupled contribution. Publishers to the Information Commons do *not* need to agree on any common ontology: the only necessary agreement is to use the common namespace provided by UUID references, and a core metadata standard to enable users to subscribe to the intended contributed content. Published content can provide any data according to any metadata standards — however, the existence of common interfaces and well-respected publishers encourages stability rather than fragmentation.

Table 2: Serialization of the u-form for *New York* as an XML / RDF file

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:cityrole="http://uform.civium.net/~01de90b7c4e2d111d689643ee15b07581d/"
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
  <rdf:Description rdf:nodeID="urn:uuid:~01fddbd808e2da11d6b9d507eb2135547d">
    <cityrole:metro_population rdf:resource="17967900"/>
    <cityrole:population rdf:resource="7262700"/>
    <cityrole:country_name rdf:resource="US"/>
    <cityrole:country_code rdf:resource="US"/>
    <cityrole:latitude rdf:resource="40.75"/>
    <cityrole:name rdf:resource="New York"/>
    <cityrole:country rdf:resource="None"/>
    <cityrole:longitude rdf:resource="-74.0"/>
  </rdf:Description>
</rdf:RDF>
```

The Information Commons and the World Wide Web

The Information Commons is readily accessible through the World Wide Web, as a source of Web Services, and as a technology for building direct web interfaces. The latter case will be described in the next section, which discusses the work of 3 Rivers Connect. The former case is discussed briefly here.

As well as the peer-to-peer network of repositories and shepherds, MAYA Design's technology also supports the provision of Web Services. That is, a repository server can also behave as a Web Service, and provide a serialized u-form in response to an HTTP request. The form of this response can be easily configured to different user needs. For example, Table 2 contains an XML / RDF file produced in response to a Web Services request from a Semantic Web agent. In this way, the Information Commons can be integrated directly into the formalism of the Semantic Web, as described by Minola and Miller [2004].

The metadata schema of this RDF file is automatically generated from the roles played by the u-form. In order to produce RDF files according to other schema, an extra u-form is created that details the mapping between schemata. This can sometimes be difficult for mathematical reasons, because (for example), the mapping of variable names from one schema to another is rarely a simple one-to-one correspondence. However, such difficulty is endemic to all attempts at metadata normalization, and is essentially a semantic rather than an engineering problem. In essence, the Information Commons can be used as a provider of Web Services with no more difficulty than any other Web Server.

However, there is a distinct advantage to be gained by using a Universal Database repository directly, which is replication. If a web interface in the NICS system relies solely on Web Services, then data will be unavailable if the remote server providing the data goes down. However, if the Universal Database technology is used as the transport layer, the u-forms in question are automatically replicated at the point of use. This means that datasets become shepherded to the machines providing interfaces to those datasets. This removes

the problem of Web Services going off-line when servers become unavailable, and has the added benefit that queries are more efficient once the information is replicated to a local venue.

3 Rivers Connect

This section describes the innovative contributions made to community information systems in Pittsburgh, where 3 Rivers Connect (3rc) has taken the lead in developing the Information Commons and using the system to bring much-needed data and services to the region.

Many community development groups, each with important social and community-oriented missions, lack the financial capacity or technological tools they need to quickly adapt to this changing economic landscape. This is where organizations such as 3rc can help. 3rc's vision is to create a regional information asset for southwestern Pennsylvania that will allow area organizations to dynamically create, combine and retrieve vital information, regardless of its original location. With these ends in mind, 3rc has embraced the ambitious mission of creating and developing the Information Commons in southwestern Pennsylvania.

By providing a flexible way to seamlessly combine data from many facets of our region, the Commons is emerging as a powerful community planning tool. It will allow leaders and organizations to understand the issues affecting our community and the resources focused on meeting needs in ways that go far beyond current capabilities. As a community of individuals and organizations, we will be able to share information, collaborate more easily, and have access to the crucial facts needed to make the best possible decisions on how to invest resources and deliver services. So far, projects using the Information Commons to benefit southwestern Pennsylvania have encompassed:

- The collection and integration of data from multiple community partners.
- A variety of simple web-based ways for community organizations and the general public to access the information.
- Applications and analytical tools to help make sense of the data, navigate to related information and draw relevant conclusions.

By simplifying the access to a combination of social, economic, and educational information sources, the Commons will ultimately help target investments with more precision (to maximize the impact of each dollar spent), improve the services delivered and allow for more informed public policy decisions. 3rc will also provide better analytical tools and visualization options (like maps), because these services and tools can be developed for the entire Information Commons, not just for single organizations and projects which could not afford such development in isolation.

Example Community Information Projects

Thanks to the financial support of the Richard King Mellon Foundation, the Heinz Endowments, the Grable Foundation and the Human Services Integration Fund, the work

of creating the Information Commons in southwestern Pennsylvania is well under way. 3 Rivers Connect has made significant progress in the last year on a number of initial projects that are important foundational 'building blocks' for the emerging regional information resource.

3rc's current projects include:

1. HumanServices.net a groundbreaking information utility of human services agency data representing a collaboration between the Allegheny County Department of Human Services and the United Way of Allegheny County. An example interface from one of the projects is shown in Figure 2.

2. Regional Education Data Aggregation (REDA). An important first step in moving toward the states vision for a unified and streamlined information management system (PIMS) and a critical aggregation tool to facilitate the inclusion of real-time K-12 education data in the Commons.

3. A+ Schools. An important community asset featuring extensive information about the after school resources available in Pittsburgh.

4. Outdoor Recreation. An exciting opportunity to integrate the online content of Venture Outdoors and the 3rc GreenPittsburgh.net site. This project represents a starting point for the aggregation of regional environmental data.

5. Content Management System. The Content Management System will provide smaller organizations in the region with a simple administrative interface into the Information Commons. It provides a web-based way to add and maintain data in the Commons.

In the first of these examples, 3rc is currently working with the Allegheny County Department of Human Services (DHS), A+ Schools and the United Way of Allegheny County to share their databases of service providers. This will increase the accuracy and coverage of services while reducing the redundancy of many agencies trying to manage the same information. From DHS caseworkers who will be able to save time and enhance the quality of their care plans to social workers in after school programs who will have real time access to public school attendance and performance records, community organizations will be able to share information more easily and collaborate to meet the needs of the children and families they serve.

The interface in Figure 2 is an example of the way the Information Commons can be used to generate direct Web access to the information collected by these projects. Users can use keyword and category search, combined with GIS mapping and navigation abilities, to find individual facilities where services are provided. These individual facilities are fused with other information to provide an integrated view as shown in Figure 3. Again, it is surprising to note that all this information — including graphics, spatial and temporal information, textual descriptions, and technical classifications — is stored in u-forms, and the Web interface is generated directly from these u-forms.

Partnerships with community organizations are an essential part of the process. Awareness and momentum are building rapidly. Through current and proposed projects, 3rc is

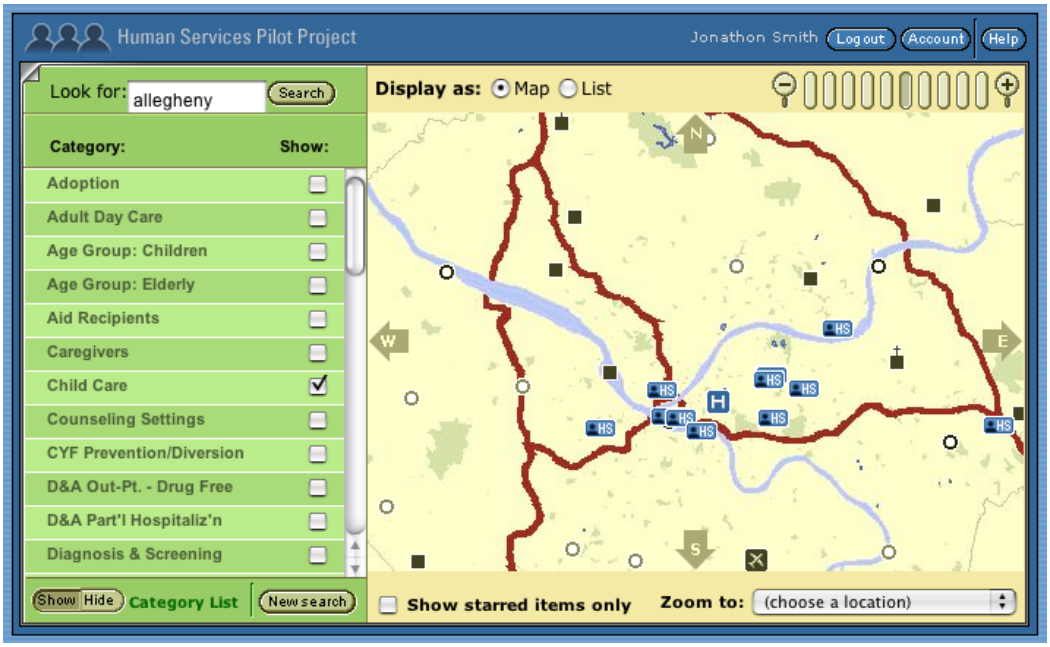

Figure 2: From the Human Services pilot project website, showing DHS facilities in the Pittsburgh Area.

Figure 3: View whose primary focus is a u-form representing an individual facility in the Human Services pilot project, with links to u-forms representing services offered, location, contact and transportation information.

working with many high profile community leaders and organizations. These include but are not limited to;

- Allegheny County Department of Human Services (DHS)
- A+ Schools: Pittsburgh's Community Alliance for Public Education
- The Brookings Institution (Urban Markets Initiative and National Infrastructure for Community Statistics (NICS) group)
- Carnegie Mellon's Center for Economic Development
- Fox Chapel School District
- Pittsburgh Partnership for Neighborhood Development (PPND)
- Port Authority of Allegheny County
- Quaker Valley School District
- State of Pennsylvania Department of Education, Community and Economic Development (DCED)
- United Way of Allegheny County

- University Center for Social and Urban Research (UCSUR) at the University of Pittsburgh
- Venture Outdoors

Linking Technologists to Practitioners

3rc's role in the growing Community Information System is critical to its success. 3rc provides the vital link between technologists and the practitioners who will actually use the system and derive benefit from it. Itself a community development non-profit, 3rc is more able to understand the information needs of other non-profits than a technology vendor would be. This is particularly important in building trust in the future. Many non-profits are naturally afraid of the concept of data liquidity. Some non-profits are afraid that if their data becomes liquid it will become mixed with lower-quality data and corrupted; others are afraid that if others can use their data, they will lose the competitive advantage they hoped to gain by investing in the data collection in the first place.

In this space, 3rc can provide a guarantee that data will be used responsibly and faithfully. Data sources will be acknowledged, and data integrity will always be preserved in the Information Commons. This is partly facilitated by the loosely-coupled architecture of the Commons: data fusion is an important service, but the sources of fused data are always preserved in *shadow* u-forms that are a faithful and intact representation of the source data.

An example of this relationship is 3rc's work with the Pittsburgh Partnership for Neighborhood Development (PPND). PPND is an active coalition of neighborhood CDCs, funding sources and financial intermediaries concerned with investment and affordable housing in Pittsburgh's low and moderate income neighborhoods. In 2003, this group, for the first time, undertook the extensive field work and reconnaissance to build a multi-dimensional parcel-based database of the property records in each of their neighborhoods. 3rc is working with them to bring this set of data into the Information Commons, where it can be combined with education, human services, environmental and census data that is already part of the commons. Once combined, this data can provide tremendous insights into the profiles of each neighborhood, map areas of need, and help local neighborhood advocates understand where environmental, social service or educational service gaps are occuring. Over the long term, this combination of datasets can lay the groundwork for robust local indicators of socio-economic progress in the region.

Conclusion

The goals for the Information Commons are ambitious. To thrive in the coming years, the nonprofit sector in southwestern PA will have to embrace structural and systemic change. Superficial fixes will not suffice. We have an opportunity to forge an exciting new path for the nonprofit agencies in our region. Our efforts are gaining statewide and national recognition for their creativity and innovation. It will take hard work, persistence and patience to make this happen. But, with the help and cooperation of many valuable partners, we are getting it done and gradually making a difference in our community.

These projects have demonstrated that it is possible to break down artificial barriers between different sorts of data. As globalization continues apace, it is ever clearer that there is a single interconnected world, and this requires a single integrated information system. At the same time, there is a growing variety of needs and points of view that must be supported by this information system.

In generating a common framework for community information, NICS is improving the way community leaders can make the right decisions, based on the best analyses of up-to-date information. Such a project needs answers for both technological and social challenges. The Universal Database technology of MAYA Design is providing the information architecture and technological infrastructure for self-describing, highly portable and flexible community information. By combining these visions in practical community-oriented projects in Southwestern Pennsylvania, 3rc is breaking new ground in enabling information to serve the best interests of the community.

References

Michael Dertouzos. *The Unfinished Revolution*. Harper Collins, 2001.

Ahmed El-Rabbany. *Introduction to GPS: The Global Positioning System*. Artech House, 2002.

P. Leach, M. Mealling, and R.Salz. A UUID URN namespace. Technical report, The Internet Society, December 2004. URL `http://www.ietf.org/internet-drafts/draft-mealling-uuid-urn-05.txt`. Current draft, awaiting approval.

Peter Lucas. Civium: A geographic information system for everyone, the information commons, and the universal database. In *Vision Plus 10*, Lech/Arlberg, Austria, 2003.

Peter Lucas and Jeff Senn. Toward the Universal Database: U-forms and the VIA Repository. Technical Report MTR02001, MAYA Design, 2002. URL `http://www.maya.com/web/what/papers/maya_universal_database.pdf`.

Peter Lucas, Jeff Senn, and Dominic Widdows. Distributed knowledge representation using universal identity and replication. Technical Report MAYA-05007, MAYA Design, 2005a.

Peter Lucas, Dominic Widdows, Joe Hughes, and William Lucas. Roles in the universal database: Data and metadata in a distributed semantic network. Technical Report MAYA-05009, MAYA Design, 2005b.

Frank Minola and Eric Miller. RDF primer. 2004. URL `http://www.w3.org/TR/2004/REC-rdf-primer-20040210/`.

Andrew D. Reamer and Pari Sabety. National infrastructure for community statistics: An overview. Working paper, Brookings Institution, Urban Markets Initiative, May 2003. URL `http://www.brookings.edu/metro/umi.htm`.

Pari Sabety and Virginia Carlson. Using information to drive change: New ways of moving markets. Report, Brookings Institution, Urban Markets Initiative, July 2004. URL `http://www.brookings.edu/metro/umi/framingpaper.htm`.