Connecting the Dots for Democratic Accountability*

Semantic Web-Based Information Sharing Policy and the Future of Investigative Reporting

By J.H. Snider, President
iSolon.org
October 22, 2010

Abstract

Since 9/11, there has been a revolution in government information sharing technology and policy for national security purposes, such as tracking potential terrorists. It’s now possible for government employees to connect the dots on national security risks across tens of thousands of government agencies at national, state, and local levels of government.

This paper addresses two central questions: 1) if the government can connect the dots for government users of national security data, why not for citizens and journalists seeking democratic accountability data? 2) To what extent is the government’s information sharing framework for national security an appropriate framework for democratic accountability?

In answer to the first question, the paper argues that the time has come to revolutionize government transparency through the use of new, governmentwide information sharing technologies. In answer to the second question, the paper argues that the government’s core information sharing standards for national security may be a good foundation for democratic accountability data standards but not the institutions created to build upon and implement those standards. It suggests that the U.S. Securities & Exchange Commission (SEC) and International Accounting Standards Board (IASB) may provide a better institutional model.

In making its argument, the paper focuses on issues associated with connecting the dots for animate objects (such as elected officials, those that influence elected officials, and those that receive resources granted by elected officials) rather than inanimate objects (such as legislation, government budgets, and campaign contributions). This is because animate objects hold political power, are the ultimate objects of democratic accountability, and are at the center of debates over fair information practices.

*This working paper was presented at “Connecting the Dots for Democratic Accountability, Washington, DC: National Press Club, October 22, 2010.

©2010 iSolon.org
Connecting the Dots for Democratic Accountability

Semantic Web-Based Information Sharing Policy and the Future of Investigative Reporting

The ability to share information among humans and across time is a defining mark of civilization, and increasing this ability has been a central driver of human progress. Historical watersheds in the ability to share information include the advent of language, writing, the printing press, and the Internet. Each of these inventions facilitated the sharing of information faster, at lower cost, across more time, and across more people. The advent of semantic web technologies, which facilitate the unambiguous sharing of information between machines, now promises another leap forward.

For this leap forward to occur in the area of democratic accountability, governments must make it easier to unambiguously identify the politically powerful agents, such as elected officials and interest groups, referenced in government databases scattered across the web. Governments are already tackling similar identifier problems in the area of national security, most notably the tracking of terrorist suspects across tens of thousands of government databases at international, national, state, and local levels of government, as well as countless private databases in industries such as airline, maritime, and rail transportation.

Following language widely adopted in the national security realm, we shall call the problem of automatically sharing unambiguous information governmentwide about politically powerful agents the “connecting-the-dots problem.” This working paper argues that the time has come for government information sharing policy to address this problem. If the government can address the problem for national security, it can also do so for democratic accountability.

On Connecting the Dots for National Security

“We propose that information be shared horizontally across new networks that transcend individual agencies.”

--9/11 Commission

“The lack of information sharing between federal, state, local authorities and the private sector, especially about individuals with ties to terror groups, was the cornerstone of the 9/11 Commission's findings. It was apparent to the 9/11 Commission’s panel members that interconnected data bits of information regarding the looming 9/11 conspiracy resided within disparate government agency databases as well as private sector data sources. The fundamental problem was not a lack of information, rather that no primary agency or decision maker had enough integrated information about fledgling terror-related conspiracies to form a mosaic-like aggregate threat picture and act to prevent them.”

--Valledor, J. C., Fort Leavenworth, KA: School of Advanced Military Studies, U.S. Army Command and General Staff College
Of course, the analogy between connecting the dots for national security and democratic accountability is imperfect. Although much can be learned by studying government information sharing technologies and policies for national security, the applicability of the analogy to democratic accountability is limited and varies depending on specific circumstances.

This paper seeks to begin a conversation about what the open government community should learn from the government’s national security information sharing initiatives. Perhaps the most important lesson is simply that governmentwide information sharing based on unambiguous tagging of data at its point of creation can be done. This paper recommends that the open government community build on the core who-what-when-where (UCore) standards framework adopted for national security, but create a significantly different institutional structure and superset of standards based on the norms of liberal democracy rather than national security. The suggested prototype for that institutional structure, based on a mix of the U.S. Securities & Exchange Commission (SEC) and International Accounting Standards Board (IASB), is only briefly touched on in this paper.

This paper can only begin the process of identifying all the questions that need to be answered before connecting the dots for democratic accountability can achieve what has already been accomplished for connecting the dots for national security. If readers come away with the conviction that the connecting the dots problem for democracy is worth solving but raises many difficult questions worthy of further investigation, this paper will have served its purpose.

**The Problem**

Representative democracy requires that voters have the information necessary to hold their representatives accountable. This requires that critical information about voters’ representatives, including who finances their campaigns and who benefits from their actions, be not only public but easily accessible.

When this information is scattered across many different databases, easy access requires that information about the

---

**On Connecting the Dots for Democratic Accountability**

“Under current law, enforcement of ethics and lobbying laws is spread widely among a variety of different agencies, which do not frequently work together, and do not provide the public information in an easily accessible format. As president, Obama will create a centralized database of lobbying reports, congressional ethics records, and campaign finance filings that is made available on the Internet in a searchable, sortable and downloadable format, and compiled and maintained by one agency and under one roof. This database will allow Americans to see easily and in one place who their legislators are meeting with; who they are collecting money from and how much money they are collecting; and to review their ethics records.”

--President Barack Obama, “Restoring Trust in Government and Improving Transparency”
same entities in different databases can be easily linked together by unambiguous identifiers. The goal of creating a semantic web, which seeks to turn the Internet into a large, searchable database, depends for the foreseeable future on the utilization of such identifiers.

Although the semantic web means different things to different people, a key concept is that data can be tagged with metadata and packages of metadata called ontologies, which facilitates the automated processing of the data by machines. When the data and metadata are both posted to the web, the web can become a giant decentralized database.

The Gary Larson cartoon in Figure 1 illustrates the basic idea of metadata and ontologies. The cartoon is funny because most children by the time they are toddlers can recognize the objects in the cartoon such as the shirt, pants, and house. They don’t have to be labeled for their meaning to be clear. A computer, however, is not so smart. But by labeling the various objects with metadata, a computer would be able to read the cartoon in the sense that it would know a shirt, pants, and house were included in the picture. Consequently, the computer could search for all pictures containing a shirt, pants, and house.

An ontology is simply a logical set of metadata. The door and window metadata, for example, would be a part of a “house” ontology.

To understand the value of metadata, consider the search for information about Tim Johnson, a member of the U.S. Senate, scattered across dozens of different databases including legislation, floor speeches, committee speeches, campaign contributions, gifts, lobbying, and earmarks. Tim Johnson is a very common name. How can one be certain when one is searching that one will only find information about the Tim Johnson who is a senator in Congress? And in the case of Tim Johnson, there are two representatives in Congress with that name, one in the House (from Illinois) and one in the Senate (from South Dakota), so one would want to be able to distinguish between the data on each of them.

Let’s say we want to use metadata to connect the dots about all the data that refer to the Tim Johnson we are searching for. In doing so, we want to avoid false positives (getting

---

On Connecting the Dots for Democratic Accountability

“We can create transparency in this modern age in which everyone who can Google can also be a research fellow for waste, fraud, and abuse in government. To do so, we have to break through a tendency in government to put up websites that are mostly propaganda while not putting up real databases that are searchable and cross-searchable against other databases. That is something that actually has been bipartisan. Chairman Towns and I have worked together to try to get data integrity that actually is searchable into every part of government. We have not yet succeeded as much as we had hoped... Chairman Towns and I put forwarded amendments to mandate... the metadata necessary for people to really know what they are looking at.

--Representative Darrell Issa, Co-chair, Congressional Transparency Caucus
results for other Tim Johnsons) and false negatives (missing data for the Tim Johnson we are searching for). If every member of Congress had a unique identifier attached to their name as metadata, then the problem of identifying all the government document references to the right Tim Johnson would become trivial. The unique identifier could read something like Congress111SenateSD1, with Congress111 for the 111th Congress between 2009-10; SenateSD for the U.S. senate seat from South Dakota; and 1 for the first senator from South Dakota.

**Figure 1. Human vs. Machine-Readable Data**

An especially common use of metadata is to make it easy for machines to track the who, what, when, and where of data. Who is acting? What type of entity is acting? When did they act? Where did they act? As we shall see, the government’s information sharing framework is built on such a simple, universal foundation.

For purposes of democratic accountability, the most important types of entities are potentially powerful political agents. These include elected representatives, agents that seek
to exert political influence on those representatives, and agents that benefit from the actions of those representatives.

References to these political agents are typically scattered in many different databases. Separate databases on members of Congress may cover such items as legislation, floor speeches, committee activities, Congressional reports, campaign contributions to leadership PACs, letters to executive agencies, dear colleague letters, meeting schedules, rooms for events on Capitol hill granted free to select interest groups, and personal office expenditures. Databases for interest groups seeking to influence legislators include campaign contribution, lobbying, gift, travel, and revolving door. Databases on large-scale recipients of government benefits include earmarks, government contracts, miscellaneous government expenditures, appointments to high profile government bodies, building permits, and business licenses.

Many of these databases may be scattered in different places, often called “data silos” in the literature on information sharing. For example, lobbying, gift, and travel reports for Congress may be stored separately in the House and Senate; earmarks are scattered on the 535 congressional office websites; and Congressional correspondence to agencies may be scattered among dozens of different agencies.

Many of these databases contain data whose potential democratic significance cannot be understood when viewed in isolation. For example, a database of campaign contributors may take on a very different meaning when it is linked to recipients of earmarks, government contracts, building permits, or licenses.

All these records are already generally considered public. But the lack of unambiguous identifiers for the same entities means that linkages between them are often difficult to make. As a result, extracting useful knowledge from the abundance of scattered databases is often not practical. That is, just as it was impractical for the law enforcement community to “connect the dots” about potential terrorists before 9/11, it is currently often impractical for those concerned with democratic accountability to connect the dots among relevant public databases. Figure 2 depicts a few of the “dots” that should be easy to connect for elected officials. Note that each dot/circle represents a type of database that may be stored in many different databases. For example, letters from members of Congress to agencies are scattered in dozens of agency offices.

A critical benefit of tagging data with metadata structured in ontologies is that such tagging facilitates automating much of the reasoning about democratic data that must now be done manually. For example, if Contributor A gives a contribution to Mayor A who appoints Commissioner X who gives a contract to Contributor A, a computer can automatically discover these indirect linkages (Snider 2009). Of course, automation is merely a tool; human judgement will still be critical to interpreting the meaning of the data.
At all levels of government, connecting the dots remains a very labor intensive and expensive process. Only at national and state levels of government do non-profits such as the Center for Responsible Politics and the National Institute for Money in State Politics even make a serious attempt to connect the dots. This is partly because there has only been foundation grant money for dot connecting at these levels of government. To connect the dots for the more than 80,000 local political jurisdictions in the U.S. is viewed as prohibitively expensive. See Sidebar A for a description of some of the problems the Center for Responsive Politics has connecting the dots (Krumholz 2010).

Even at the national level of government, it is only affordable to integrate the most important and easy to access databases. And within each of those databases, the focus often has to be on only the most important data, such as the largest contributors. Databases with very valuable metadata about individuals and organizations, such as the U.S. Securities & Exchange Commission (SEC; information about public company ownership), United States Patent & Trademark Office (USPTO; information about the trademarks under which organizations do business), state level offices responsible for corporate registrations, local licensing boards (approximately one fourth of the U.S. work force is licensed by government), and U.S. Department of Treasury (information about financial institutions) are
either ignored or accessed indirectly via expensive proprietary sources, such as Hoover’s, that compile the government data. Linking between inputs (such as campaign expenditures) and outputs (such as government contracts) is rarely done except for high profile applications such as earmarks. Data aggregation for elected representatives, such as members of Congress, is much more comprehensive than for appointed officials, such as commissioners appointed to agencies, despite the fact that much more public information is available for the latter than the former.

Local and state governments often have more complicated democratic structures than the federal government. At a state level, for example, the executive power might be divided into many different elected offices (such as Secretary of State, Attorney General, School Superintendent, and Comptroller). At a local level, in addition to dozens of elected offices (e.g., judges, orphan’s court, and registrar of wills), appointed members of public bodies (e.g., school board, library board, and planning commission) may be granted substantial discretion to allocate resources and be nominally subject to the same or more rigorous public disclosure than the elected officials who appointed them. Nevertheless, with rare exceptions, all the data exist in separate silos and remain unintegrated.

Just as employers have a right to track how their employees spend their money, citizens in a democracy generally have a right to track how their elected officials

### Sidebar A

#### Sample identity resolution problems

**ORGANIZATIONAL IDENTIFIERS**

**Recycling Names:** Chevron → ChevronTexaco →
Chevron; Time Warner merges with AOL →
AOL Time Warner → Time Warner.

**PR Name Changes:** Philip Morris → Altria, but
Philip Morris is retained as a *subsidiary* of
Altria.

**Same Name, Different Organization:**

- state/regional/local banks with generic names such as Citizens Bank (e.g., enter “Citizens Bank” in Google and more than a dozen banks with that name will show up in the search results; MCI refers to a big telecom company and scores of local companies as well, only some of which are subsidiaries of MCI).
- **In Name Only:** The rights to a brand name may be sold to another company without the business that goes along with it (e.g., in 2004 IBM sold its PC business to Lenovo granting the right to use its brand name for PCs for up to five years).
- **Joint Ventures:** Sony + Ericsson → SonyEricsson;
Dow Chemicals + Corning → Dow Corning.

**Complicated Histories:** Travelers merges with Primerica which merges with Citi, then sold to St Pauls, which becomes Travelers.

**INDIVIDUAL IDENTIFIERS**

**Common Names:** Is this Robert Smith that Robert Smith?

**Missing Suffixes:** Are the two Bob Jones at the same address one a senior (Sr) and another a junior (Jr)?

**Same Person, Different Addresses:** Is Mr. Rich, with four homes in four different states, and writing checks from each of those locations, the same person?

**Same person, different occupation:** Did Bill Lee switch jobs or is this a different Bill Lee?

**Complex Family Relationships:** Who is a spouse when there are divorces and remarriages? (E.g., Aby Rosen married Samantha Boardman Rosen, previously married to Elizabeth Wechsler Rosen but separated in 2000 and divorced in 2004.)

Many large families have family members that work in the same company with lots of different spouses and other close relatives.

Source: Interview with Sheila Krumholz, October 2010.
spend their money. But tracking recipients of government funds is often far harder than is generally understood. As a general rule, only the easiest problems, such as tracking the recipients of earmarks or contracts, have been addressed. By “easy” is only meant doable. Those that attempt to tackle such easy problems rarely find them anything but “easy.”

Consider the seemingly easy task of finding out how much an individual or average K12 public school teacher in a particular school district is paid. Despite the fact that data purporting to provide this information are widely distributed and reported, it is next to impossible to do and not in fact done.

Admittedly, most states have laws mandating that teacher compensation be publicly disclosed. But those laws don’t address the connecting-the-dots problem, with the consequence that finding out how much teachers are actually paid is all but impossible.

In this author’s public school district, for example, teachers receive a base salary based on a grid with step increases for years worked and education credits obtained. They also receive many salary incentives, including bonuses for working in Title I schools, teaching certain subjects, securing a national teacher certification, cashing in unused sick and personal days, supervising after school activities, and otherwise working beyond a 38.5 hour work week and 189 day work year. In addition, they receive medical, dental, vision, education, and other benefits, some of which are for retirement rather than current use. In theory, at least, the school district must disclose the individual components of their compensation, but it is not obliged to make it easy to connect all the dots.

Moreover, teachers now also receive a significant amount of compensation from sources outside the school district, including dozens of different government agencies. Some teachers in this author’s school district, for example, receive extra compensation from county, state, and federal government agencies, as well as private education schools, for work done in the classroom during regular school hours. This compensation may include reimbursement for college tuition as a reward for becoming a teacher, extra compensation for working in poor schools and in hard-to-staff subject areas, subsidized workforce housing, tax credits for school supplies, extra compensation for national teacher certification, extra compensation for supervising an ed school student in the classroom, and in-state state pension contributions. Consequently, with current financial reporting systems, it is all-but-impossible to connect the dots to find out how much such government employees are actually paid.
In addition to an improved ability to connect the dots, governmentwide democratic data standards should be understood to create huge economies of scale in both data production and consumption. We have seen the benefit of data standardization in many other walks of life, including weights & measures, electrical plugs & voltages, roads, train tracks, airline communications, shipping containers, product bar codes, and financial accounting. The success of the Internet, too, has depended on browsers utilizing basic standards such as http, html, and css. Standards for sharing democratic data governmentwide promise similar huge gains in productivity, including lower costs of data entry (due to factors such as less rekeying of duplicate data and less need for employee training in the use of obscure data formats and technologies), lower application costs (because the same applications can be used across thousands of databases), and more full featured and diverse applications (because the market for applications would be orders of magnitude larger than it is today). In Sidebar B, the U.S. Program Manager for the Information Sharing Environment describes some such cost efficiencies for disaster assistance (Kshemendra 2010, 52).

**Why Focus on Humans vs. Things**

The need for widely adopted standards to facilitate the unambiguous identification of the same agents across multiple databases is part of a much larger need to create semantically enriched data for democratic accountability. This includes semantically enriched budget data, legislative data, and conflict-of-interest data (Snider 2010, 2010,

---

**Sidebar B**

**Case Study:**

**Efficiency gains from sharing data in disaster assistance**

“The Disaster Assistance Improvement Program (DAIP) is the result of citizen confusion and frustration with conflicting and misdirected information received as they sought disaster assistance from multiple Federal programs in the aftermath of Hurricane Katrina. The goal of DAIP is to ease the burden of disaster victims by consolidating federally funded forms of assistance information, application intake and status information into a unified system....

This information exchange is high-volume, particularly in the wake of national disasters when individuals require access to application assistance. The DAIP information exchange reduces the time necessary to identify and apply for Forms of Assistance (FOA)—instead of completing multiple forms, applicants can submit one assistance form which can be used for other disaster relief requests. This information exchange streamlines the disaster assistance application process, leading to increased satisfaction in verifying the status of and receiving loan assistance.

The DAIP information exchange also reduces data entry of application information and reduces time needed to respond to application status inquiries. DAIP enables information to be exchanged across many organizations, including federal, state and local disaster relief agencies such as FEMA and SBA’s Office of Disaster Assistance. Additionally, the DAIP information exchange promotes the reusability and consistency of information by processing applications with fewer user input errors.”
The public, for example, should be able to easily compare school budget data across many criteria across the 14,000+ public school districts in the U.S.; it should be able to easily search for legislation on particular subjects across the 99 state legislative bodies in the U.S.; and it should be able to easily search for linkages between campaign contributors and recipients of government expenditures.

A well designed and implemented set of democratic data standards could increase the efficiency of searching for certain domain specific information by a factor of a million or more. For example, good government groups now link conflict of interest data to earmark data because this is relatively easy to do (see Figure 3). The earmark data is disclosed by each member of Congress and rarely is more than a few hundred items per member. In contrast, this author is not aware of any good government group or media organization that currently links conflict-of-interest data to budget data (see Figure 4). The reason is related to the fact that even a small state or a large city may have millions of transactions that serve as the basis of its budget. Linking the ambiguous vendors to the ambiguous contributors would be an overwhelming task. But if the original data had been entered in such a way as to make the direct or indirect links unambiguous, making the linkages could be as fast and affordable as doing a Google search. And once the linkages were automated, countless new ways of displaying the data could also be automated.

Figure 3. Linking conflict-of-interest to earmark data

For U.S. Sen. Richard Shelby, the largest earmark recipient for fiscal year 2009

One reason to focus on the unambiguous identification of agents (e.g., people and organizations) is because they are the objects of democratic accountability. Tracking inanimate objects (e.g., political jurisdictions, financial statements, and legislation) may be very important, but they are only a means to track animate objects, and it is animate objects, not inanimate ones, who are sensitive to being tracked and wield political clout.
Consequently, tracking animate objects raises unique and controversial concerns, such as the protection of civil liberties and trade secrets. Whereas standards for tracking inanimate objects can often be delegated to obscure expert bodies, standards for tracking animate objects need to be subject to extensive public deliberation.

Information sharing standards for the unambiguous identification of agents also serve as a foundation for the information sharing standards of more specific domains of knowledge, such as the above-mentioned domains of legislation, budgets, and conflicts of interest (the latter domain usually placed under the heading of “ethics”). This helps explain why the government’s information sharing initiative for tracking terrorists is built upon a core set of agent identification standards that is shared by all the domain-specific information standards such as those for justice, public safety, emergency, disaster management, intelligence, and homeland security (U.S. Department of Justice's Office of Justice Programs 2008, p. 7; Kshemendra 2010).

The Timeliness of the Problem

Policies to enhance the transparency of government information can be hierarchically organized. As lower level transparency problems are solved, higher level ones come to the fore. We can distinguish between four public record eras based on the type of technology used:
In the file cabinet era, public records are stored in physical file cabinets. A large fraction of public records, especially in local government, are still stored in file cabinets.

In the computer era, public records are stored in computers only accessible on local networks physically located within an agency. Most public records at all levels of government are still probably in this era, which has contributed to the call to make it possible to “Google your government.”

In the Internet era, public records are freely accessible over the Internet without the need for a password. This is the fastest growing type of public records access, but only a small fraction of public records are currently available in this fashion.

In the semantic web era, public records are enhanced with metadata that make it easy to connect the dots when the public records are scattered across the Internet. The metadata can either be added by the producers or the consumers of the data. But it is generally agreed that it is far more efficient and effective for the producers than the consumers of data to add the metadata, so this is how enterprise wide data sharing systems, such as connecting the dots for national security or customer management systems, are usually designed.

Of course, there aren’t sharp distinctions between these eras; they are only useful heuristics. For example, parts of public records systems, such as legislative information systems, may be computerized (e.g., the tracking of bills) whereas other parts (e.g., the tracking of committee votes) may be in file cabinets. Computer data may be gradually opened up over larger and larger networks until they are opened up to the public Internet (there is also the common situation of posting ostensibly public records on the Internet but requiring a password to see them). And data posted to the Internet may be gradually supplemented with more and more metadata that are more and more widely used across different government entities.

One important transition stage between the Internet and Semantic Web eras is the posting of structured data to the Internet (Robinson et al. 2008). This is a great improvement over the early Internet era because structured data is a type of metadata. A government agency, for example, provides much more useful data to the public when the data are in a structured Excel spreadsheet than an unstructured pdf. This allows for vastly more efficient mashups and other manipulations of the data. But the structured data is also far more valuable to the public if it is structured using a standard format that allows it to be searchable across multiple databases. This is the difference between a single legislature such as the U.S. House of Representatives sharing its legislative information in a downloadable XML file (the XML
tags are the metadata that explain the meaning of each structured field) and using a widely adopted legislative ontology to structure the data. It is noteworthy that African legislatures have adopted such an ontology, whereas American legislators send out press releases when they merely decide to publish structured data unique to their own legislature (Snider 2010, 2009).

Another important transition stage is posting data with standardized metadata and posting the data with standardized metadata to the open Internet. This is the difference, for example, between using XML tags and RDF tags. XML tags allow information sharing across computer networks (often called “interoperability” in the computer literature) but not necessarily the open Internet. RDF tags, an extension of XML, expose metadata to the public Internet. From this perspective, XML tags are a semantic technology for data sharing, but RDF tags a more powerful semantic technology for data sharing. The federal government’s connecting-the-dots initiatives are currently primarily based on XML.

Public records problems need to be solved hierarchically; that is, on a first-things-first principle. If you’re in the file cabinet era, the highest priority is to get to the computer era. If you’re in the computer era, the highest priority is to get to the Internet era. Only if you’re already in the Internet era, does the highest priority shift to get to the Semantic Web era. Since we are rapidly entering the Internet era for public records, the need to shift focus to the emerging Semantic Web era is becoming imperative.

Consider this author’s experience searching for interest group data in the mid-1990s when he was researching his Ph.D dissertation. He was trying to investigate all the different ways the National Association of Broadcasters exerted political influence on members of Congress and other elected officials. The federal data he sought were scattered in seven different databases: the campaign contributions database (at the Federal Election Commission) and the separate lobbying, travel, and gift disclosure databases (at the U.S. House of Representatives and U.S. Senate). Only one of the databases (campaign contributions) was online. And each database had to be searched separately.

In the course of his research, he discovered that the National Association of Broadcasters (NAB) was coded with many different letter sequences including National Assoc. of Broadcasters, National Assoc. Broadcasters, National Broadcasters, and NAB. Even within the same database, such as lobbying disclosure, the identifiers for the NAB varied. Perhaps the problem was the result of printed forms that didn’t include enough space for long trade association names, or database fields that were cut short to save on computer memory, or filers who used shorthand simply to save time in filling out voluminous paperwork, or filers who were more than happy to obscure their tracks. For a large trade association such as the NAB that employed dozens of lobbying firms, each of which had many clients and had to file separately, there were clearly many opportunities for inconsistency.
But the lack of unambiguous identifiers was not his major search problem. It was transporting himself to Washington, DC from Chicago and visiting the House and Senate offices that housed the data. Once he had already made that huge commitment of resources, it was relatively simple for him to use his common sense to know that the “National Association of Broadcasters” and “NAB” referred to the same entity.

Moreover, he didn’t even consider searching for the comparable information about the NAB’s lobbying activities at the state and local levels of government. The time and cost of transporting himself to all those locales rendered it out of the question.

Today, more of those databases are online and increasingly in a structured, downloadable format. The barrier to more efficient and effective public access, therefore, is increasingly the difficulty in searching online across the various databases.

For national security, this problem is now vividly understood. After terrorists destroyed the World Trade towers on September 11, 2001, the 9/11 Commission studied the information available to the government at the time of the attack (National Commission on Terrorist Attacks upon the United States. 2004). It determined that various government databases already had the information necessary to prevent the attack but the difficulty in sharing the data—connecting-the-dots—led that information not to be used effectively. Thus, connecting data about potential terrorists across tens of thousands of international, national, state, and local government entities, as well as across many critical private sector industries, became one of the top priorities in the battle against terrorism (Valledor 2010). Subsequently, the sharing of national security data was extended far beyond the identification of potential terrorists (Kshemendra 2010).

So why not extend information sharing to the full range of democratic accountability data? That is, in what ways are the government’s information sharing technologies and policies for national security a useful model for democratic accountability?

The Connecting-the-Dots Model for National Security
The government’s efforts to connect the dots for national security across governments and private enterprises are stunningly ambitious. They involve sharing data across tens of thousands of local government entities, such as police and fire departments; many separate state and federal agencies, such as the Department of Justice, Department of Homeland Security, Department of Defense, and Department of Transportation; foreign countries; and countless private sector entities, such as the reservation systems of airline, train, and boat companies.

The governmentwide model for connecting the dots for national security can be divided into three major components: Technical standards for sharing national security data, institutions
to implement those standards across government entities, and disclosure norms and incentives.

**Technical Standards**

The technical standards for information sharing are largely incorporated in the National Information Exchange Model (NIEM). NIEM provides a common language with which federal, state, local, and tribal agencies can describe, structure, and share information. As Paul Kshemendra, the Program Manager for the U.S. Information Sharing Environment formulates the need for such a language in his 2010 Annual Report to Congress: “Without a common lexicon—a lingua franca that all participants can understand—meaningful information exchange is impossible (Kshemendra 2010, 47).” NIEM is designed to facilitate exchange of information across a family of different domains such as justice, public safety, emergency, disaster management, intelligence, and homeland security (U.S. Department of Justice's Office of Justice Programs 2008, p. 7; Kshemendra 2010). Figure 5 depicts the basic structure of NIEM, including its flexibility to support many different domains of knowledge, each with its own specialized information sharing standards.

**Figure 5. NIEM supports many different domains of knowledge**
Another way to view NIEM is as a hierarchy, with the universal core of standards for information sharing at the base. For purposes of connecting the dots for democratic accountability, it is the universal standards, UCore (available at www.ucore.gov), that are most relevant because they are designed to be universally applicable to data sharing and are a suitable foundation for data sharing for both national security and democratic accountability. In contrast, NIEM’s higher level domain specific information sharing standards, such as those specifically for data sharing within the U.S. Department of Justice, are less relevant.

In theory, domain specific information sharing standards for democratic accountability, such as for legislation or conflicts-of-interest, could be added to the NIEM framework. Indeed, NIEM has already developed a set of information sharing standards to track government stimulus funds (see www.recovery.gov). But for the most part, NIEM does not currently incorporate such standards or have any plans to do so.

Since UCore is potentially very important for connecting the dots for democratic accountability, some of its technical details are described below. UCore covers the most commonly agreed upon and universal concepts of Who, What, When, Where. It is designed to be simple to understand, explain, and implement. Who, for example, distinguishes between a person and an organization and includes simple properties to identify people and organizations such as alternate name, citizenship, and contact information. Where describes a location using geolocation (e.g., longitude and latitude), a physical address (e.g., a postal address), and a cyber address (e.g., a URL). Time describes either an instant in time or an interval of time. And What describes the type of Who such as member of Congress, registered lobbyist, or public corporation. See Figure 6 for a simple data model of these four concepts.

UCore is not intended as a complete solution to information sharing in any specific domain of knowledge. Its aim is to provide a lot of functionality with minimal effort.

Even with these very simple who, what, where, when concepts, powerful geospatial analysis (analyzing activities via geographic maps), temporal analysis (analyzing activities over time), and link analysis (analyzing relationships between entities) can be greatly enhanced.

One prominent technology writer, Michael Daconta, has called UCore “the Twitter of information sharing” because of the information sharing power it “The intention of [UCore] is ambitious: Craft a universal core of the most common data elements across all possible exchanges. By definition, that universal core lies at the center of all possible domain intersections. The question is: What are the most common things everyone must agree on to have minimal interoperability? The answer embodied in the standard is refreshingly minimalist: who, what, when and where.”

--Michael Daconta, “UCore’s Giant Leap”
It packs so concisely and efficiently (Daconta 2009; see also Daconta 2008).

**Figure 6. The Four Basic Concepts of UCore**

The same author describes UCore as part of the paradigm shift from a producer-centric to consumer-centric information sharing mindset (2009). Adding UCore metadata upfront adds to the burden on producers of data but lightens the burden on consumers of data. It’s a one-time upfront cost that can dramatically lower the collective cost of consumers to access the information.

If you’re an open government advocate or investigative reporter, the temptation is to make do with the data that government producers have generated and supplement it with your own metadata to facilitate search. The long-term result of making do with consumer generated metadata is a far more expensive dot connecting process that generates far less democratic accountability.

There is surely value in having data consumers supplement the metadata provided by governments. But most of what is currently done by government data consumers should be done by government data producers because the metadata are a public good and can be added to data far more efficiently and effectively during data production.
Institutions

The federal government has developed a complex institutional structure to manage governmentwide information sharing for national security. The structure is in part hierarchical, with the federal government spending billions of dollars to help motivate government agencies at all levels of government to share their data. The various functional data silos, such as police, fire, justice, intelligence, and national defense also have their own internal hierarchical structures to facilitate information sharing. At the top, many of the efforts are monitored and coordinated by the Program Manager for the Information Sharing Environment (PM-ISE) located in the Executive Office of the President of the United States. The PM-ISE co-chairs the White House-based Information Sharing and Access Interagency Policy Committee (ISA IPC).

Unlike the technical standard UCore, this institutional structure is not suitable for extension into the domain of connecting the dots for democratic accountability. This is because the normative framework for national security and democracy are too different. Following this line of reasoning, the National Information Exchange Model (NIEM) could be renamed the National Security Information Exchange Model (keeping the same acronym, NIEM) while creating a new model, the Democratic Accountability Information Exchange Model (DIEM). Both would share UCore but would otherwise be separate. Figure 7 illustrates the proposed relationship between UCore, NIEM, and DIEM.

Figure 7. Linking NIEM and DIEM via UCore

DIEM should be embedded in an institutional structure with substantial autonomy from the legislative and executive branches. This could be fostered by creating an Open Government
Commission (OGC) modeled on existing federal commission structures such as the Federal Trade Commission (FTC), Federal Communications Commission (FCC), and Securities & Exchange Commission (SEC).

A commission is run like a mini-legislature and made up of commissioners appointed for a fixed length of time. In an executive agency, in contrast, a single individual runs the agency and can be fired or otherwise removed from office at the pleasure of a higher level public official.

As with federal appointees, the U.S. president nominates commissioners for approval by the U.S. Senate. But commissioners are usually allocated by political party. The Federal Election Commission (FEC) has an equal number of commissioners from each party. The SEC, FCC, and FTC each have one more member from the majority party. Another design is to have an equal number of commissioners from each party and have them jointly appoint an independent, tie-breaking commissioner. If the two parties cannot agree on the independent commissioner, then a coin flip may be used to choose among the Republican and Democrat nominees, an independent panel of judges may make the choice, or some other independent selection mechanism may be used.

The legislative like operation of a commission, the terms of fixed duration for commissioners, and bipartisan power structure, make a commission a relatively independent and checks & balances type of democratic institution. Ideally, an Open Government Commission (OGC) would be granted even more autonomy than a typical commission.

To create an even greater degree of independence, OGC could farm out standards setting to an independent body, the Democratic Accountability Standards Board (DASB). As analogies, the SEC farms out creating accounting standards to the Financial Accounting Standards Board (FASB), and the European Commission and most developed countries to the International Accounting Standards Board (IASB).

The IASB is especially noteworthy because it began as a think tank and continues to be a primarily private entity. Yet it has become the most influential financial accounting standards setter in the world. Moreover, it sets the financial standards for dozens of independent and highly developed countries (Veron 2007).

The OGC would have the power to create rulemakings to issue democratic metadata standards and collect statistics about their use. Its rulemakings would follow the strictest implementations of the Administrative Procedures Act, which would make its standards setting process far more visible than that associated with NIEM. Its statistics generating capacity could be analogous to that of the Bureau of Labor Statistics.

More controversially, OGC could have some type of mild enforcement power, such as the ability to withhold federal disbursements for up to six months to states and localities that were
in gross violation of its data standards. The theory here is that no state or local government
deserves federal monies unless it has done a minimal amount to allow citizens to monitor it
and thus hold it accountable. The federal government has often used a democratic
accountability standard in its disbursements of funds to foreign governments. It has also
demanded controls from state and local governments to ensure that its funds are efficiently
and effectively spent. The U.S. Department of Education, for example, requires standardized
local disclosure of student test scores as a condition for receiving federal funding. Many
federal grants are even conditioned on a requirement for at least some democratic
deliberation. An example is the requirement that local public bodies hold public meetings
and provide public notice of them to discuss how grant monies should be disbersed. The
rationale is that such a public process ensures that the grants are dispersed according to
public need. These democratic controls need to be modernized and focused more on data
standards.

Conditioning federal grants on the implementation of minimum democratic accountability
data standards could also save the federal government money on oversight, especially of
grants to state and local governments. This is partly because it would reduce the cost of
direct federal oversight. But probably more important, it would greatly increase the ability of
the public (including investigative reporters acting on behalf of the public) to do more of the
federal government’s oversight. The federal government could thus more confidently rely on
being alerted by the public if the grants weren’t being administered in response to public
need. This is similar to the logic behind the transparency requirements for federal agencies
included in the Administrative Procedures Act (McCubbins, Noll, and Weingast 1987). It is
also part of the rationale for standardized formats for state and local governments to report on
the use of Federal stimulus monies, public school student test scores, and environmental
variables such as clean air. But the application of this economic logic, especially for state
and local reporting, has been haphazard.

**Disclosure Norms and Incentives**

Some of the obstacles to information sharing for national security and democracy are similar.
Others are very different.

Information sharing for national security and democracy both need to follow a basic set of
“fair information practices.” The Center for Democracy and Technology provides the
following useful framework for fair information practices (Dempsey February 2007).

1. Notice (or openness) -- the government should state when it is collecting data,
through a published notice and wherever possible on an individual basis - Privacy
Act, subsection (e)(2), (3) and (4). Sometimes notice is discussed in conjunction with
choice or consent, the principle that individuals should have control over when data is
collected and how it is used, unless a certain standard is met for the compulsory collection of data (such as with a warrant or subpoena).

2. Purpose specification -- the government should specify the purpose for which it is collecting data - Privacy Act, subsection (e)(3).

3. Collection limitation – collect no more than relevant and necessary for the specified purpose (minimization) – see Privacy Act subsections (e)(2) and (7).

4. Retention limitation – retain (or "maintain") no longer than necessary for the specified purpose – the Privacy Act does not require record disposal, but see subsections (e)(1), (2) and (5).

5. Use and disclosure limitation – limits on secondary use without consent – Privacy Act subsection (b).

6. Data quality – timely, accurate, complete – Privacy Act subsection (e)(5).


8. Access to one’s own records – Sometimes referred to as “individual participation,” access is a right in and of itself, but also it is a precursor to exercising the right to insist on the correction of mistakes – Privacy Act subsections (c), (d) and (f).

9. Redress – Sometimes combined with the “individual participation” principle, this refers to the right to challenge inaccurate data, preferably before adverse decisions are made, and to correct mistakes and obtain redress for abuse – Privacy Act subsections (e) and (d).

10. Accountability– audit, enforcement – Privacy Act subsections (e)(9) and (10), (g) and (i).

However, implementation issues for national security and democratic accountability differ significantly. These differences drive the need to create a different institutional model for democratic accountability data. Important differences in disclosure norms and incentives include the following:

**National Security data tends to be secret; democratic accountability data public.**

Intelligence data for the most part must be kept secret. If suspects know what data is available about them, they can take countermeasures to avoid detection, capture, and punishment. In contrast, democracy is based on the principle of political equality, which entails equal access to information for all citizens (Dahl 1989). For example, the official actions of elected officials are generally expected to be public and accessible for all citizens, not just insiders. The notion that only insiders would be allowed to track the official actions of public officials is generally anathema to the democratic ideal.

**Subjects of national security data tend to have fewer means to view and correct faulty data.** Because national security data tends to be secret, it is harder for its subjects to monitor and correct faulty data. On the other hand, because democratic accountability data is more public, the reputational harm from faulty data may sometimes be greater.
National security data users tend to be government employees, not investigative journalists and other private citizens. Partly as a consequence of the secrecy associated with national security data, authorized users must often have government security clearances and are usually government employees. In contrast, it is desirable that private citizens be the primary users of democratic accountability data. The government must screen, hire, and train national security analysts, but since no one wants or expects the government to screen, hire, and train investigative journalists, democratic accountability data also costs the government less to use.

Agents identified in national security data tend to have greater means and incentive to hide their identities. We expect terrorists to go out of their way to obfuscate their identities and have many means at their disposal to do so. Elected officials, interest groups, and recipients of government expenditures have less incentive and means to obfuscate their identities. This is hugely important because it means that it may be practical to put the onus on agents to correctly identify themselves with a unique (but anonymous, see below) identifier. Nobody knows their own identity better than themselves. If sources of democratic accountability data can be trusted to unambiguously identify themselves, the cost of generating unambiguous identifiers for them can drop by orders of magnitude (Jonas 2007).

National security databases need to collect more contextual information about agents in order to uniquely identify them. Since terrorists and other suspects are unlikely to accurately identify themselves, national security databases tend to gather as much contextual information as possible to unambiguously identify subjects. This includes such information as a suspect’s home address, other addresses, birthday, education, work, car, boat, travel itinerary, and just about any other information that might help establish the unique identity of an agent. But if democratic accountability databases could trust agents to accurately identify themselves, less contextual data about agents would need to be collected. Unambiguous identifiers, by reducing the amount of information about agents that needs to be collected and made public, thus can enhance privacy.

The vocabulary to describe agents in national security databases tends to be more pejorative than the language appropriate for a democratic accountability database. Phrases such as “watchlists,” “suspects,” “workups,” and “encounter data” are inappropriate to describe information sharing in a democratic accountability context. Different, less pejorative language needs to be developed to describe democratic accountability data.

The politically powerful have more to lose from democratic than national security information sharing. It is both human nature and rational for power seeking individuals to want as little as accountability related information about themselves as possible to be made publicly available. Information is power, and insiders who have an information advantage over outsiders like to keep it, which is why lobbyists often prefer to operate in the dark (Birnbaum 1992) and why legislative information systems in the U.S. are so primitive.
In a democracy, the principle of political equality entails that all political actors have as equal access as possible to information about their elected officials, including officials’ actions and potentially significant conflicts of interests. Elected officials and powerful interest groups can therefore be expected to oppose the development of an efficient and effective democratic accountability sharing environment whereas they would presumably have no such personal objection to development of other types of sharing environments, such as those tracking terrorists or sex offenders. As a result, control over democratic accountability data standards must be as independent as possible from the direct influence of elected officials and powerful special interest groups.

**Reciprocity as an incentive to share data tends to be a more powerful motivator for national security than democratic accountability.** When government agencies seek to track an entity such as a potential terrorist or a motorist’s driving record, they have a common goal and can gain mutual advantage in sharing information as the price necessary to receive information. To get, they must give. It is less clear that stewards of democratic accountability data can gain similar advantages by facilitating data sharing. Reciprocity works best when free riding is impossible. An unfortunate secondary consequence of the public nature of democratic data is that it makes free riding relatively easy.

**Connecting the dots for national security requires expensive, government controlled indices.** Since national security databases tend to be secret, the government must take on the role and expense of aggregator; that is, providing the links or index to the various databases. Since democratic accountability databases are public, aggregation can be left to competitive private entities such as Google, Yahoo, and Bing, or more specialized aggregators that focus on democratic accountability data, such as the Center for Responsive Politics, Maplight, and the National Institute on Money in State Politics.

**Concerns about protecting trade secrets are less important for national security data.** Partly because national security databases tend to be secret, preservation of trade secrets tends to be of less concern. As a general rule, companies will not want competitors to have easier access to government held information about themselves. Consider building permits, which need to be public for democratic accountability purposes, if only because access to favorable building permits drives so much of local politics, including who contributes to political candidates and lobbies elected officials. Many companies would like to retain as much control as possible over when competitors can get access to their building plans because building plans can reveal valuable information about their future plans. However, it should be reflected upon that private data aggregators such as Choicepoint, Acxiom, and Zillow.com already compile a huge amount of public information about both individuals and organizations (Ayres 2007, 134–8). If much of this information is already available and affordable for companies to purchase, the harm from making it more broadly available to average citizens seems relatively small. Although concerns about trade secrets are often
overblown, they have been given short shrift in the literature on fair information practices and need to be thoughtfully balanced against the need for more democratic accountability.

**Linking Technology & Fair Information Practices**

New technology to connect the dots fundamentally changes old assumptions and practices concerning the appropriate balance between public disclosure and privacy.

As a general rule, democracy requires more disclosure about powerful than weak political agents. Thus, we have disclosure thresholds on what level and in how much detail contributions need to be disclosed. But these thresholds were created in a technological environment where connecting the dots was prohibitively costly. As that environment changes, the nature and the amounts of the thresholds should also be reconsidered so that they better reflect democratic principles.

An especially important technological development may have to do with the ability to anonymize personal identifiers (Jonas and Harper 2010, 319). The new anonymization techniques may reframe many of the old conflicts between disclosure and privacy. They may make it possible, for example, to connect the dots while retaining privacy—a difficult concept to grasp within the current technological framework for democratic disclosure.

It may be feasible, for example, to irreversibly anonymize an identifier, add up all the campaign contributions associated with that identifier, and then only make the original identifier public if a certain threshold level of campaign contributions is reached.

The old system of attaching addresses to campaign contributors may also no longer be needed if the underlying purpose of the address is to identify the various political districts associated with the identifier. If an address could flexibly be transformed into its associated political districts, it is possible that the address itself would no longer have to be made public as part of campaign finance and other political disclosure requirements.

**Setting Priorities**

Creating an information sharing environment for democratic data is likely to take decades and move much more slowly than creating an information sharing environment for national security. The focus, therefore, should be on prioritizing the most realistic short-term applications, while also planning for the long-term. Applications should be prioritized based on the least controversy, least cost, and greatest impact.

**Controversy**

The least controversial applications should be the initial focus for creating and implementing democratic accountability data standards. For example, disclosing data on organizations is less controversial than on individuals, so the first priority should be on organizations. Among individuals, large contributors and recipients of public expenditures should be prioritized.
over small ones because it is less controversial to disclose information about the most powerful political agents. Of course, controversy is largely in the eye of the beholder. Collecting data about elected officials is not likely to be controversial for the general public, but will likely generate fierce opposition from elected officials (Snider 2009).

**Cost**

Less costly applications should be preferred. For example, public data already near the top of the public record hierarchy, such as data already posted on the Internet, should be given priority over data near the bottom, such as data still locked in file cabinets. Similarly, small public data sets, such as earmarks, should be given priority over large public data sets such as transaction level budget data.

**Impact**

Applications with greater impact should be preferred. For example, campaign disclosure databases generally contain more useful information than gift disclosure databases, so they should be given priority. Similarly, budget data is generally more useful than building permit data, so should be given a higher priority.

Unfortunately, there is very little correlation between the controversy, cost, and impact of making a given set of democratic accountability data shareable. The best strategy may be to focus on the democratic data associated with the least political controversy, which may mean starting with the data least useful for democratic accountability.

Focusing on data about the executive branch is probably much more practical than focusing on data about the legislative branch, because the legislative branch makes the laws and has relatively little incentive to give up control of information about itself. This helps explain why the executive branch consistently outpaces the legislative branch in implementing open government principles. The current focus on using standardized data for government budget disclosure is probably as non-controversial a major use of data sharing technology as is possible.

Institutionally, the initial focus has to be on working through existing institutions such as the U.S. Office of Management and Budget, the White House’s Office of Science and Technology Policy, and the Congressional committees with jurisdiction over government reform.

However, this must be balanced by an active private effort to develop and adopt data sharing standards for democratic accountability. A precedent is the private development of FedSpending.org, which tracks government contracts and served as the basis for the development of the government’s USAspending.gov. However, the best analogy may be the International Accounting Standards Board (IASB), which started as a private think tank and has subsequently come to dominate the worldwide development of government mandated
financial accounting standards. ICANN, OASIS, and W3C are other private standards setting bodies whose standards have been adopted by government.

Partly because elected and other public officials have less of an intrinsic desire to make themselves rather than others, such as terrorists and sex offenders, accountable, grantmaking foundations and non-profits need to play a larger role in the effort to connect the dots for democratic accountability than they did in the effort to connect the dots for national security. However, it should be noted that the Markle Foundation played an important role in that effort, which could serve as a model (e.g., Markle Foundation Task Force on National Security in the Information Age March 2009).

Convincing private democratic accountability data aggregators to adopt information sharing standards for their most important data may be a practical short-term strategy. Representatives of such groups should meet together to develop a sharing strategy, perhaps starting with a subset of UCore, even something as simple as its data standards for when and where.

Government agencies not usually included in democratic reformers’ agendas, such as the SEC, USPTO, and state level agencies that incorporate organizations, need to be added to those agendas because they hold critical democratic agent identifier information necessary for effective information sharing.

Conclusion

In recent years, the technology to automate the sharing of data between databases has undergone dramatic improvement. This is reflected in the government’s massive effort to connect the dots governmentwide for national security in the wake of 9/11. No similar broad ranging effort to connect the dots governmentwide has occurred with democratic accountability data. This raises the question: if the government can facilitate connecting the dots for national security purposes, why not for democratic accountability purposes, too? Why does America have to continue to live in the dark ages when it comes to shining a light on the official actions of its elected officials?

The primary lesson to be learned from efforts to connect the dots for national security purposes is simply that governmentwide data sharing across tens of thousands of separate government entities, each with their own databases, is indeed possible. That is, where there is a will to share government data, there is a way to share it. But is there a will to share such data? That may be the million dollar question.
The UCore standard developed to share national security data could also serve as a foundation for sharing democratic accountability data. However, the governmentwide institutions created to connect the dots for national security are not well suited to develop and implement democratic accountability data sharing standards because of the fundamentally different norms and incentives associated with national security and democratic accountability data. For example, connecting the dots for national security tends to include non-public information, whereas connecting the dots for democratic accountability tends to include public information.

Consequently, a fundamentally different institutional framework should be created to foster governmentwide information sharing for democratic accountability. A central feature of this institutional framework is that it should reflect a checks & balances rather than executive agency model of accountability. The SEC and IASB, rather than the White House’s information sharing czar, the PM-I$E$, or its Office of Science and Technology Policy, may be useful models for this purpose.

If this paper launches a discussion of how to move American democracy into a new governmentwide data sharing era, it will have served its purpose. This author hopes to follow up with a half day or full day conference on these matters in late spring 2011 or early fall 2011.
References


Jonas, Jeff. 2007. To Know Semantic Reconciliation is to Love Semantic Reconciliation. In *Jeff Jonas*.


