Title
Improving Federal Spending Transparency: Lessons Drawn from Recovery.gov

Authors
Yee, Raymond
Kansa, Eric C
Wilde, Erik

Publication Date
2010-05-25

Improving Federal Spending Transparency: Lessons Drawn from Recovery.gov

Raymond Yee, Eric C. Kansa, and Erik Wilde
School of Information, UC Berkeley

UC Berkeley School of Information Report 2010-040
May 2010

Available at http://escholarship.org/uc/item/7tw2w9wx

Abstract

Information about federal spending can affect national priorities and government processes, having impacts on society that few other data sources can rival. However, building effective open government and transparency mechanisms holds a host of technical, conceptual, and organizational challenges. To help guide development and deployment of future federal spending transparency systems, this paper explores the effectiveness of accountability measures deployed for the *American Recovery and Reinvestment Act of 2009* ("Recovery Act" or "ARRA"). The Recovery Act provides an excellent case study to better understand the general requirements for designing and deploying "Open Government" systems. In this document, we show specific examples of how problems in data quality, service design, and systems architecture limit the effectiveness of ARRA's promised transparency. We also highlight organizational and incentive issues that impede transparency, and point to design processes as well as general architectural principles needed to better realize the goals advanced by open government advocates.

Contents

1 Introduction

In this document we present some general principles for making the Federal Budget and spending more intelligible and available for public review and debate. The significance of budgetary transparency is difficult to overstate. Information about federal spending can affect national priorities and government processes, having impacts across our society that few other data sources can rival. Budgetary transparency is therefore central to the whole program of "Open Government." Because the information systems that support budgetary disclosures are so important, they must be carefully designed to offer effective and meaningful forms of transparency.

Key aspects for enabling budgetary transparency include:

- *Open Machine-readable Data in Non-proprietary Formats.* Information should be published in ways that maximize the potential for reuse, aggregation, and analysis on a variety of computing platforms. Non-proprietary formats are essential because they enable consumers of information to use open source or commercial computing applications of their choice. Non-proprietary formats also increase the longevity of information since these formats can be best supported by archival systems.

- *Identifiers to Linked Data.* Budgetary datasets will reference many different codes and coding systems. Unfortunately, entities identified by these codes are often difficult to find. If they do exist on a public website, explanation for codes often exist in difficult to parse PDF documents. Such explanatory data should be published in more reusable formats and services that provide contextual information and help make budgetary data intelligible. Ideally, URIs (*Uniform Resource Identifiers* [2], most commonly HTTP Web links) should be the main identifiers used in budgetary disclosures. URIs can identify linked Web resources that contain important information that give context and meaning to budgetary data. Linked data will make budgetary data easier to understand by clarifying the rationale and purpose behind budgetary appropriations and expenditures. Linking data will also enable agency officials and lawmakers to better gauge program performance and effectiveness.

- *Services for Dynamic Data.* Many "Open Government" and "Open Data" advocates will want to see bulk-download features for budgetary data. While bulk-download features are important and should be implemented, we stress that bulk-downloads are *not sufficient*. Bulk-downloads are most suitable for relatively static datasets. For the Federal Budget, we anticipate that different federal agencies and other government monitoring offices will be continually releasing relevant data. Some of this information will include ongoing spending as new expenditures, grants, and contracts are authorized. In other cases, data may be continually collected on program metrics and reports. New programs may be announced (along with identifiers for these new programs), and other programs may close down. Thus, the information landscape of budgetary information, including key data required to make sense of the Federal Budget will be constantly evolving. Since bulk-download approaches typically deliver data well "after the fact" they must be complemented by other approaches that offer more real-time access. Because of the dynamic nature of budget disclosures, Web services (such as those we proposed at `http://recovery.berkeley.edu/` [8, 9]) need to be implemented. Web services make it easier to incrementally retrieve relevant data, retrieve and monitor updated data, and if well designed, make these data easier to use on a wider variety of platforms. Thus, Web services not only facilitate independent analysis, visualization, and interpretation of budgetary data, but they also enable these independent efforts to be conducted in real-time, based on continually streaming data.

- *Data Archiving and Versioning.* Building trust is a critical element in making transparency work. The public needs to trust that information published today will not be gone tomorrow. Data archiving and curation services are an essential component to transparency. The public must also see guarantees that information will not be altered without notification or explanation. Update and notification services

are essential, as well as version tracking, so that the public can see the history of changes on sometimes critical data. Furthermore, budgetary information should be citable; that is, different versions of budgetary disclosures should be clearly identified and reliably retrievable into the future.

2 Case Study: American Recovery and Reinvestment Act Transparency Measures

The accountability measures around the *American Recovery and Reinvestment Act of 2009* ("Recovery Act" or "ARRA") provide an excellent case study of the interplay of the general principles outlined above with transparency effectiveness. We can glean many lessons for short and long term visions of federal spending transparency from Recovery.gov.

An *Office of Management and Budget (OMB)* memo (April 6, 2010) providing guidelines for the immediate future of federal spending transparency explicitly cites ARRA fund transparency, stating that the memo builds on "achievements of and the lessons learned from implementing the American Recovery and Reinvestment Act" [10]. Recovery.gov and Federalreporting.gov are likely models for what we will see in future systems, especially because policy makers have singled out these systems (especially Recovery.gov) as exemplars. Transparency advocates should therefore pay close attention to how these disclosure systems work in practice. By examining these real world systems, we can learn about advantages and problems of different design approaches.

Specifically, we can look at ARRA transparency measures through the criteria specified in the OMB memo: *accessibility, completeness, accuracy, and usability*. In each of these critical dimensions, we should have expectations for the Recovery Act. President Obama made transparency a centerpiece of this legislation promising that Americans would be able "to see how every penny in this plan [the Recovery Act] is being spent."[1] Such absolute precision may be beyond the capability of any accounting system, especially given ARRA's vast scale and complexity. However, we can examine Recovery.gov and associated systems to see whether we can gain a reasonably intelligible and comprehensive understanding of ARRA spending. Given this more reasonable expectation for ARRA transparency, how effective is Recovery.gov in promoting public understanding and accountability?

3 Missing End-to-end Transparency: Guesswork and the ARRA Jigsaw Puzzle

To begin with, a fundamental requirement for budgetary transparency is the ability to relate actual spending with the intentions of Congress as expressed in legislation. Ideally, we should be able to trace *the full lifecycle of federal spending*, from the agency's budget request to congressional appropriation to disbursement by federal agencies. For example, we should be able to easily track fund allocations from the Congressional record, which documents various versions of the bill, to the final enrolled bill PL-115.[2] that the Congressional Budget Office (CBO) scored (the source of the $787 vs $792 billion figure). Once the bill becomes law, we should then pick up the trail for how the Treasury implemented ARRA provisions and allocated money to different accounts, as identified by *Treasury Account Symbols (TAS)*. From these accounts, we should see how spending unfolded across various agencies and then to the primary and secondary recipients. At each stage, we should have a clear tally of the total number of all relevant entities (e.g., recipients, awards, treasury

[1] "Organizing for America — OFA Blog: Message from President Obama: 'What recovery means for you'." http://my.barackobama.com/page/community/post/obamaforamerica/gGxHGc (Accessed May 13, 2010)

[2] "Public Law 111 — 5 — American Recovery and Reinvestment Act of 2009." http://www.gpo.gov/fdsys/pkg/PLAW-111publ5/content-detail.html (Accessed May 13, 2010)

accounts) and expected sums of funding allocation and spending, as well as an account of the data sources and reasons for any uncertainty.

Unfortunately, reality deviates from this ideal of end-to-end transparency. ARRA data is more like a jigsaw puzzle than a clear tabulation and accounting of spending. Members of the public must invest great effort in making even partial sense of the whole picture, and this process involves a great deal of guess work. Some difficulties should be expected since ARRA, by its very nature, is a complex and sprawling piece of legislation. Investing transparency measures dynamically "on the fly" while coordinating across so many players will naturally lead to gaps and problems. However, these gaps and problems persist and continue to stymie ARRA transparency. It is simply too difficult to know with any real confidence how "every penny in this plan is being spent."

These difficulties should give policy makers pause when they look to the ARRA as a model for making the overall picture of the federal budget and federal spending more transparent. Below we describe some of the specific problems and gaps in ARRA transparency and some strategies to improve transparency effectiveness.

3.1 PL 111-5 and Making Sense of the Legislative Process

Ideally, open government transparency measures should offer the public channels to understand, evaluate, and impact key democratic processes. Understanding the legislative process and how the legislation gets implemented in the Executive branch of government should be a major high level goal. Unfortunately, ARRA transparency measures were scoped too narrowly and largely ignored the legislative process.

Current Problems

- *Need for Machine-Readable Legislation:* The run-up to the final piece of legislation was very confusing; every financial analysis used a different way of catgorizing key budget items. It is very difficult to understand spending levels proposed in different versions of the legislation and how these ultimately became law. One could compare the House and Senate versions of the Recovery Act vs the final piece of legislation, though it is a non-trivial task to do a detailed comparison. There was a big effort by news organizations such as the *New York Times*[3] and *ProPublica*[4] while the final details of the bill were being passed to parse the bill. Since the bills were not in machine-readable form, journalists had to manually transcribe numbers into spreadsheets and then come up with separate tallies. It was next to impossible to get a definitive analysis and compare the various analyses to see whether they actually agreed and where they differed.

- *Need Transparency in CBO Analyses:* Moreover, the CBO released preliminary and final analyses of ARRA in PDF form but not a detailed analytic breakdown on how it arrived at the numbers (including simple matters like adding up all the line items and showing how they do in fact add up to what the analyses came up with). In other words, the CBO analyses are effectively a black box, whose accuracy is difficult to assess and therefore challenge or correct. This is especially problematic in fast moving situations like the passage of ARRA.

The overall point of this discussion is to highlight how transparency in government spending is not just an issue for the executive branch of government; it also places requirements on legislative transparency. Ideally, transparency should promote accountability. Through greater visibility in how laws are created, debated, and enacted, reformers can better identify weakness in political and government processes and build support for improvements.

[3]Hossain, Farhana, Amanda Cox, John McGrath, and Stephan Weitberg. n.d. "The Stimulus Plan: How to Spend $787 Billion — The New York Times." http://projects.nytimes.com/44th_president/stimulus (Accessed May 13, 2010)

[4]Grabell, Michael, and Christopher Weaver. n.d. "The Stimulus Plan: A Detailed List of Spending — ProPublica." http://www.propublica.org/special/the-stimulus-plan-a-detailed-list-of-spending (Accessed May 13, 2010)

Proposed Solutions:

1. *Clear Identifiers Needed for Legislative Line Items:* Funding provisions in all bills should carry clear and unambiguous identifiers expressed in machine-readable formats. These identifiers can enable tracking of funding items through CBO analytical processes and later into the Treasury's accounting systems.

2. *Open Data for CBO Analyses:* The supporting data behind CBO analysis should be made public along with the final report. Datasets showing how line-items actually add up would be very helpful, and these can be linked back to specific provisions in bills to show their context in large and complex bodies of legislation.

3. *Open Access to CBO Models and Analytic Formulas:* Since budgetary analysis represents much more than adding up several line items, the CBO needs to provide formulas and financial assumptions, as well. For example, the tax implications of ARRA spread over ten years and can involve complex computations. Sharing these analytic tools can lead to their refinement overtime, especially if it becomes easier to relate actual budgetary data back to CBO projections.

3.2 Linking Data: Identifiers, End-to-End Tracking, and Comprehensiveness

Just as it is difficult to trace the spending estimates used in the legislative process, it is nearly impossible to know how actual spending patterns mapped to Congress's wishes as expressed in the passed legislation. How do spending provisions in law lead to allocations of money in specific programs, accounts, and projects? How are legislative decisions implemented across various agencies and administrative divisions?

Current Problems

- *Missing Connections between Legislation and Executive Processes:* Through some effort, one can look at items in the original legislation and see how they may fund different programs (TASes) in ARRA. However, there is no official and unambiguous way to tie a given line item in PL 115 to a specific award flowing from that item.

- *Lack of Comprehensive Catalogs of Identifiers:* There is no definitive, public list of Recovery-related TAS identifiers. [5] Without this list it is impossible to get basic understanding of how the Treasury responded to this legislation. It is also impossible to understand the degree to which Recovery.gov data represents the complete and comprehensive picture. We have no way of knowing what should be expected and what data may be missing.

The degree to which reported data represents a complete and comprehensive account of spending must be communicated and demonstrated. As discussed above, there is no definitive list of *all* TAS identifiers used in the Recovery act. Without this list, it is impossible to know whether recipient reporting is 20%, 50%, or 90% complete and comprehensive. Moreover, the Recovery Act authorized exceptions on detailed reporting. In how many instances did such exceptions apply? How much money flowed through channels that did not require detailed reports?

Proposed Solutions:

1. *Publish Machine-Readable Catalogs of Identifiers:* There are many units of analysis and other entities that can directly or indirectly help explain ARRA reporting data. These include data about recipients,

[5] Yee, Raymond. 2009. "ARRA Treasury Account Symbols: the outcome of our FOIA request." Data Unbound. http://blog.dataunbound.com/2009/11/23/foia-outcome/ (Accessed May 13, 2010)

relationships among recipients (contractor-subcontractor or subsidiary relations), federal agencies and bureaus involved, and information about administrative divisions (states, municipalities, congressional districts, counties). Authoritative services to share updated lists of these identifiers would give a reliable basis for understanding if reporting measures gave complete and comprehensive accounts of government spending. The catalogs should also be available in machine-readable formats to facilitate software applications.

2. *Define URI/URL Templates:* Identification systems and coding systems need to be easy to resolve on the Web. Well defined templates for looking up Web resources for an entity associated with a specific identifier or a code improves the usefulness of coding systems. The templates themselves should be easy to read and understand for a layperson, and in doing so, they will be easier to understand for third-party software developers. To improve the longevity of Web identifiers, URI/URL templates should also avoid dependencies on the specifics of back-end implementations such as scripting languages (such that ".php," ".aspx," ".net" should not be in Web addresses).[6]

3. *Enable Linking of Data Across Agencies:* Much budgetary data only makes sense through reference to entities that are not directly described in a given dataset. The use of clear identifiers (ideally Web URIs) can provide key information to help link administrative processes within these various sectors of the Executive Branch with the wishes of Congress as expressed in final legislation. Spending data on different programs can be better contextualized with the use of Web identifiers. We should be able to follow links from spending data to see documentation about a program, relevant metrics reporting on the goals and achievements of that program, and who is responsible for overseeing that program. We also should be able to follow links from spending disclosures on a specific program back to the original legislation authorizing funding for that program.

3.3 Data Consistency, Quality, and Data Integrity Enforcement

Current Problems Measures to enforce data quality and consistency are critical to making data intelligible and usable. This is especially true given the scale and complexity of Federal spending. For the ARRA, there are several data quality problems that have hampered the utility of data and have cast doubt on the overall reliability and trustworthiness of disclosure methods.

- *Confusing Terminology, Unevenly Applied:* Transparency measures will fail if no effort goes toward making basic concepts and data intelligible to the public. Educational aids need to be offered so the public gain grasp the basics of how federal budgeting and spending works. For example, public discussion of ARRA referenced two different grand totals, $782 billion and $797 billion, numbers differing by $15 billion dollars. One estimate came from the CBO as a "budget authority" number, while the other came from the same CBO estimate as an "estimated outlay." Yet the rationale and meaning of these two different numbers remains obscure to the public.

- *Lack of Controlled Vocabularies:* Cities are fields in ARRA recipient reporting. Unfortunately, the public datasets have poor-quality city data. For example, there are numerous misspellings of "Los Angeles" making it difficult to classify and summarize financial reports by cities.

- *Key Data Missing or Omitted:* In many cases, ARRA recipient reporting has critical gaps that short-circuit understanding. In some examples, there are reports of sub-recipients without a corresponding primary recipient. In other cases, there are award keys with more than one primary recipient. In other examples, the local amounts add up to more than the award amount. Sometimes the TAS is only in

[6] For a comprehensive discussion, see: Berners-Lee, Tim. 1998. "Hypertext Style: Cool URIs don't change." http://www.w3.org/Provider/Style/URI (Accessed May 13, 2010)

the form of a placeholder ("91-XXXX"). These missing data can be found in many recovery reports without explanation.

- *No Notification of Updates:* Citizens can download the recipient data from http://www.recovery.gov/FAQ/Pages/DownloadCenter.aspx, but it is very difficult to tell what version of the data is currently available for downloading. There are no services (such as Atom feeds) for sharing update notices. There is no provision for obtaining archived earlier versions of data.

To mitigate some of these problems, spending reporting systems must take a variety of measures. The list below is not exhaustive, but represents some basic approaches to tackling some of the critical quality and integrity issues that can harm transparency efforts.

Proposed Solutions:

1. *Use Existing Controlled Vocabularies:* Reporting systems need to make better use of controlled vocabularies in data entry. The US Census already produces a list of cities, which should be reused in this context. Enforcing a controlled vocabulary for congressional districts would have prevented the political embarrassment over non-existent congressional districts. Use of controlled vocabularies should not be too rigid, since it is sometimes important to accommodate fuzziness and "edge cases." In such cases it is better to have an "other" option, so that there is some assurance that the responder is stating explicitly he or she is opting out of the choices and not making an entry error. A catalog of controlled vocabularies that is readily available and understandable to implementers of federal spending applications would help tremendously.

2. *Enforce Data Model Constraints:* Since Recovery.gov and/or Federalreporting.gov creates an award_key identifier, it can certainly ensure that not more than one primary recipient can be associated with that key. The existence of award keys with more than one primary recipient calls into question the integrity of data collection efforts.

3. *Reuse Existing Data to Promote Quality:* One would assume (perhaps incorrectly) that Federalreporting.gov should know what recipients should be reporting. After all, awards to the primary recipients are coming from the agencies. These agencies can give Federalreporting.gov an authoritative list of recipients. Hence, if a sub-recipient reports to Federalreporting.gov, that sub-recipient should be able to choose from a pre-defined list of awards. Using such measures, there should not be an award_key without a primary recipient.

4. *Business-Rules for Error Checking:* Invite the community to collaborate on a data diagnosis dashboard. The *Recovery Act Transparency Board (RATB)* and public interests groups can define certain business rules that should be in place to test data quality (e.g., one-to-one correspondence between award_key and primary recipient, award_amount must equal total of all local_amount, etc). These tests can help improve quality of data entry and monitor the quality of data once reported.

5. *Report Errors and Version Data:* The open source software development community has well developed practices for identifying problems and publicly track issues. An issue repository or analogous system can help officials, public interest groups, and members of the public identify problems, track progress on fixes, and propose solutions.

6. *Data Citation:* A critical need for making Recovery.gov more trusted and reliable is to make its content citable. One needs solid assurances that the expected version of a dataset will be obtained at an expected Web URI. These kinds of measures complement community versioning and issue tracking discussed above.

3.4 Summary: How Data Silos Hamper ARRA Transparency

One of the largest and most systemic obstacles to greater government transparency comes from the "siloed" mentality of many agency officials and government contractors. "Data silos" are systems defined by their inability to share common types of information. When key data are trapped in individual systems, inconsistencies emerge across these different systems. It becomes much harder to coordinate key government processes, assess performance, or account to the public.

The problems of data silos and the administrative dysfunctions that lead to such silos can be seen in ARRA reporting and disclosure systems. Many of the shortcomings in ARRA transparency come from lack of coordination across different sources and types of government information. As discussed, there is no way to simply connect the intent of the original legislation and the actual spending reports of ARRA funding recipients. Data about the Congressional record and ARRA simply lack the unambiguous identifiers (or "hooks") needed to establish these clear ties.

The lack of apparent coordination between the Treasury and OMB exemplifies our concerns over silos. We have yet to find an authoritative crosswalk between TAS/OMB coding systems. We were able to obtain such data through a *Freedom of Information Act (FOIA)* request, through collations developed by ProPublica, and in some "easter-eggs" inadvertently hidden in Excel spreadsheets provided by Recovery.gov. Obviously, this sleuth work to obtain incomplete "guesstimates" of how TAS amd OMB track money does not represent an ideal picture of transparency. Since OMB and the Treasury both maintain key identifiers needed to track and understand budgetary processes across the Federal Government, it is absolutely critical that they offer a publicly accessible system to reconcile their coding systems.

Similarly, there are telling discrepancies and problems in coordinating information across different sources such as *USAspending.gov* and *Recovery.gov*. Doing this alignment should be easy but is non-trivial because of subtle ways in which DUNS numbers (a unique identifier developed by Dun & Bradstreet and used by the federal government to identify award recipients) and award keys that are supposed to be stable identifiers are not actually always the same between systems. Bridging across these systems is important, because with comparison with USAspending.gov, it is difficult to understand how patterns in Recovery.gov spending compare with the overall picture of the Federal budget. In addition, USAspending.gov offers APIs (services) giving updated streams of data, while Recovery.gov offers bulk downloads. Ideally, we should have both of these capabilities in both systems.

Without specific policy efforts to work against data silos, the problems of reconciling data from different systems will only get replicated. In the future, Congress may enact additional legislation like ARRA that will require specialized websites analogous to Recovery.gov for reporting. However, the key issues for all such specialized sites will be how they relate to the larger context of federal information, including other budgetary disclosure systems such as USAspending.gov.

4 Lessons from ARRA: Agencies Need Incentives to Publish Useful Data

The above discussion of ARRA highlighted several shortcomings, many of which relate to systemic problems of "data silos" in government contexts. This point illustrates how technology choices take place in organizational and political contexts. Budgetary transparency measures will fail if agencies do not have a vested interest in opening data. Without clear incentives, officials will make poor and constraining technology choices, and the quality of implementations will suffer. Currently, officials see little reason to look beyond their own agency needs. This helps motivate the creation of "silos" that trap data within a limited and narrow range of interfaces.

Agencies need to be rewarded for providing data, via reliable services, that see reuse, aggregation, and integration with other government and citizen applications. Agencies must also be responsive to requests

from members of the public and from other parts of the government to make data available in formats and services that maximize convenience and uptake. Hiring a contractor that holds a few conference calls among transparency advocates does not demonstrate meaningful commitment to public collaboration. Consultation processes need to have real impact. Outside agencies and the public interest groups need ways to articulate specific requirements and evaluate deployments to guide revisions that better meet public needs. In other words, meaningful consultation, issue tracking and improvements need to be continual, not just at the initial planning stages. To effectively harness consultation processes to design and build systems, agencies should adopt participatory design methodologies. If agencies saw clear rewards and incentives for providing data and services that see reuse, they will invest the effort and thought needed to support interoperability.

The contracting process, including the *Request for Proposals (RFP)* and the Smartronix bid for building Recovery.gov was less than ideal for building public trust in ARRA transparency. The RFP saw little public input or consultation, and the Smartronix bid itself was only released in a heavily redacted form. These government procurement procedures do not lead to optimal outcomes, and the problems we see in Recovery.gov stem from the deficiencies in the development process. Thus, the process of building transparency systems should be regarded as important as the outcome. Collaborative and iterative processes of design help build the trust, accountability, and shared problem solving that transparency advocates seek. Therefore, participatory design methods should be considered essential for building transparency systems.

5 Citation, Trust, Curation, and Integrity

In bridging across silos, agencies need to consider measures to make their data more reliably identified and permanent. Citation is absolutely integral to understanding. It enables and expresses collaborative knowledge production across space and time, and it is the foundation on which evidence and arguments are identified, assembled, reused, and critiqued. This is true in both formal professional settings of academic communication, and in public debates expressed online in weblogs, forums, and social networks. On the Web, citations usually take the form of a hyperlink (URI/URL) that identifies a specific information resource and, while doing so, makes its retrieval fast and efficient.

For budgetary transparency, citation issues represent a key concern. As already discussed, citable, linked data is required to help provide needed context to budgetary data. However, citation also involves issues of reliability and trust, making data version control and archiving critical issues. These requirements represent key challenges, especially since important information for making sense of the Federal Budget will be distributed across many different agencies and data systems. To help address these issues, designers of systems that support budgetary transparency need to consult the digital library community. This community has developed extensive expertise in designing technologies, business processes, and policies to safeguard data and its integrity, identification, and reliable retrieval [1]. The organization *Citability.org*[7] also provides invaluable guidance for making Web based government information more trustworthy and easy to reference. This group has explored important issues about granularity and specificity in citation and retrieval.

In some ways, the citability and long term curation of different versions of datasets may represent a more important priority than "signing data." Large and complex datasets will likely contain some errors, even with good data validation practices. If we required agency officials to "sign data" as completely accurate, we may create incentives to withhold key information for fear that their accuracy will never be 100% perfect. Instead, we should reward agencies that make their data cleansing and error correction processes more open and transparent. Version tracking and citation can help foster a more collaborative and less adversarial relationship than some signing proposals.

[7] http://citability.org

6 Getting the Big Picture: Metrics, Classification and Discovery

Classification will underpin many uses for budget data. Datasets will need different forms of administrative, technical, and descriptive metadata to be managed. Given the scale and complexity of the federal government, budgetary metadata requirements will likely be similarly complex. The *National Information Exchange Model (NIEM)* may provide a good foundation for meeting these government-wide standards needs. The *Extensible Business Reporting Language (XBRL)* represents another potentially applicable standard that may be useful for describing many financial transactions in government.[8]

Standards and classification concerns are not only technical matters; they are also important issues impacting the usability of data. Searching and retrieving relevant information from large bodies of complex data is a challenge for many information services. Keyword searches are common solutions to this problem. However, keyword searches often yield incomplete and ambiguous results. The "hit or miss" nature of keyword searches can limit the effectiveness of transparency. Poor search capabilities can lead to omissions of relevant data. In turn, this can reduce trust in a disclosure system. Information retrieval systems need to be more robust, reliable, and predictable if they are to support transparency objectives and earn public trust.

To avoid some of the difficulties associated with keyword searches, metadata standards can be used to support "faceted navigation" systems appropriate for exploring and retrieving spending data. In faceted navigation applications, users can leverage sophisticated classification systems using simple and intuitive "point and click" selections. Users progressively home in on more specific information from a larger collection. Because filters are applied across an entire collection, users have greater certainty in the comprehensiveness of their results than with keyword searches. Feedback, in the form of subtotals for the numbers of items that fall under each available facet, guides users in the selection of additional filters [6]. This feature offers users important information cues about the size and composition of the collection they are searching. Dynamic feedback also helps users understand sometimes very sophisticated classification systems.

While classification can be an important enabler to understanding, classifying information across the vast Federal bureaucracy is no easy matter. It can and likely will be debated and contested. Classification exists in social and organizational contexts. Some types of classification may work well in one context but may work poorly in another context. Moreover, classification is not necessarily an objective process. It is shaped by the assumptions and goals of people and organizations. These worldviews and goals often see disagreement and evolve over time.

Thus, not all metadata, including performance metrics, will be appropriate in all situations. Failing to pay adequate attention to context can be harmful, especially with top-down imposed schemes. For example, without careful alignment to an organization's mission, poorly applied metrics may create incentives for the wrong outcomes. Agencies, programs, and the public need to come together to shape classification systems so that they reflect both the need to understand general government-wide processes and the more particular concerns of a specific program or agency. Similarly, metadata systems need to be responsive to evolve to better meet both existing needs and those not yet anticipated. While there is a clear need for government-wide standards, efforts to make the Federal Budget intelligible must also accommodate the great diversity of program goals, needs, and contexts. Just as certain metrics need to be carefully shaped to be appropriate for a given context, classification and description of budgetary data must also allow for enough nuance to facilitate and not obscure understanding.

[8] XBRL is an open data standard only in terms of access to the equivalent of its 'source code'. However, the governance structure of the XBRL consortium differs greatly from open source approaches. Paid membership and a focus on transacting business at physical conferences deviate from open source practices, and represents severe barriers to entry in participating in XBRL standards development. Because of these issues, it may be better not to use only XBRL, and consider NIEM and other standards as well.

7 Looking to the Future: Web Technologies

It is important for transparency measures to be open and easily used by the broadest constituents possible. This overarching need should guide decision-making on various implementation styles and technologies. However, many constituencies on the Web have different opinions on how best to deploy Open Government systems. Different technical styles will be promoted by different vendors, academic experts, and other stakeholders. Agency officials and systems architects will no doubt face many different sales pitches and expertly informed opinions. For example, proponents of the Semantic Web [3] (or "Linked Open Data", which is another term for the same set of technologies) often promote a specific suite of technologies centered on the *Resource Description Framework (RDF)*. RDF tools and technologies can be both powerful and very sophisticated. They enjoy widespread interest in many academic, research, and some commercial communities. In other cases, enterprise computing vendors may promote their specific vision of "Service Oriented Architectures" (usually highly planned and elaborate distributed computing frameworks). On the other hand, developers oriented toward the public Web often favor more "document-centric" and/or *Plain Web* [7] technologies. These other constituents typically favor a different suite of technologies, including Atom feeds and other RESTful services, together with XML [4] and JSON [5] for expressing structured data. Each technical style has its own set of proponents, advantages, and disadvantages, and the landscape constantly shifts as Web technologies continually evolve.

The current Web sees many different approaches working in parallel to meet different needs and solve different problems. Good architectural decisions, especially with regard to Web identifiers (URIs), will enable the Federal Government to best serve the needs of multiple communities, including both Semantic Web and Plain Web proponents. Ideally, budgetary disclosure systems should not favor one contested approach over another or pick individual technology winners and losers. Government transparency efforts therefore should not emphasize one technical style at the expense of another, but should expose information to multiple kinds of interactions. Agency officials should evaluate systems design proposals on the basis of how well they support multiple approaches to using data. Participatory and collaborative design methodologies that involve users (including third-party programmers) and other stakeholders will help agency officials prioritize technology choices that meet actual needs.

8 Conclusions

Policymakers need to understand that the key problems in making government data more open and useful come mainly from organizational challenges and less from technology. Without concerted attention on incentives, administrative processes, and institutional concerns, technologies such as "Service Oriented Architectures," the "Semantic Web," or "Cloud Computing" will remain empty buzzwords rather than real design solutions. A single system or technological choice will not automatically yield meaningful forms of transparency. Rather, officials need to identify strategies to better align agency priorities toward greater openness and collaboration between each other and the public. By institutionalizing and rewarding collaboration, information access, and data portability, agency officials will have the right processes and incentives to find appropriate technologies that best support greater transparency.

References

[1] Micah Altman and Gary King. A Proposed Standard for the Scholarly Citation of Quantitative Data. *D-Lib Magazine*, 13(3/4), March 2007.

[2] Tim Berners-Lee, Roy Thomas Fielding, and Larry Masinter. Uniform Resource Identifier (URI): Generic Syntax. Internet RFC 3986, January 2005.

[3] Tim Berners-Lee, James A. Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, 284(5):34–43, May 2001.

[4] Tim Bray, Jean Paoli, C. Michael Sperberg-McQueen, Eve Maler, and François Yergeau. Extensible Markup Language (XML) 1.0 (Fifth Edition). World Wide Web Consortium, Recommendation REC-xml-20081126, November 2008.

[5] Douglas Crockford. The application/json Media Type for JavaScript Object Notation (JSON). Internet RFC 4627, July 2006.

[6] Marti A. Hearst. Clustering versus Faceted Categories for Information Exploration. *Communications of the ACM*, 49(4):59–61, April 2006.

[7] Erik Wilde. The Plain Web. In *First International Workshop on Understanding Web Evolution (WebEvolve2008)*, pages 79–83, Beijing, China, April 2008.

[8] Erik Wilde, Eric Kansa, and Raymond Yee. Proposed Guideline Clarifications for American Recovery and Reinvestment Act of 2009. Technical Report 2009-029, School of Information, UC Berkeley, Berkeley, California, March 2009.

[9] Erik Wilde, Eric Kansa, and Raymond Yee. Web Services for Recovery.gov. Technical Report 2009-035, School of Information, UC Berkeley, Berkeley, California, October 2009.

[10] Jeffrey D. Zients. Memorandum for Senior Accountable Officials Over the Quality of Federal Spending Information. Office of Management and Budget Memorandum, April 2010.