DATE: Nov. 24, 2008
TO: John Leibovitz, Tech Transition Team
FROM: Barath Raghavan, UCSD
SUBJECT: `USAspending.gov` and the State of Government Data Transparency

For one year `USAspending.gov` has exposed $18 trillion in federal spending, yet in that year not a single dollar tracked by the site has been spotlighted by politicians or the media. While the transparency it provides is a positive step, its neglect stands in sharp contrast to the praise it initially received from watchdog groups and voters alike. The message is clear: sunlight is an insufficiently strong disinfectant; government data transparency for its own sake does not automatically improve oversight by ordinary citizens. For that data to be useful in monitoring the actions of government, users need **context**—context that is lacking today—and they need access to the **data** itself in myriad forms rather than to a predefined service to access that data. I believe that a confused notion of what transparency itself *is* causes this gap between the theory and the practice of open government. Below, I unearth contracts from `USAspending.gov` to help illustrate how these problems manifest themselves today and to help identify how we might rethink government transparency.[1]

1. In 2007 a business named Presidential Airways, Inc. received $28 million in contracts. A few months later, the name of the business had changed to EP Investments, LLC. `USAspending.gov` had little more about either, except that they provided "Passenger Air Charter Service." Subsequent searches at other non-government sites uncovered choice information about these businesses: I found that Presidential Airways, Inc. is a sister company to Aviation Worldwide Services, LLC and both are owned, perhaps indirectly, by Blackwater Worldwide Inc. through its subsidiary Blackwater Aviation.[2] The database is rife with such business metamorphosis: VECO (involved in bribery in Alaska, is now part of CH2M Hill and received $2 billion in 2006), MZM Inc. (involved in bribery in CA-50, first morphed into Athena Innovative Solutions and then was purchased by CACI International Inc., itself a recipient of $1.5 billion per year), and General Atomics (involved in suspect lobbying via representatives of CA-41, received both contracts *and* substantial *grants*).

2. In 2003, Tom Delay (TX-22) was elected house majority leader. That year, his district received $515 million in contracts; one year later, it jumped to $1.357 billion and again to $3.375 billion in 2005. The data indicates that roughly 80% of the spending in TX-22 in 2005 was for contracts that had no competition.

3. After 2000, spending on dam operation dropped sharply—from $10 million in 2000 to just $1 million in 2001–2003 and $2 million in 2004. The opposite trend is apparent in paper-shredding contracts, with $452 thousand on paper shredding in 2000 that grew to $3.5 million per year by 2007. A perhaps more ideological trend is visible in spending on abstinence education programs, which grew from $0 per year in 2000-2003 to $87 million in 2008.

The above contracts highlight the two most serious issues with the federal spending transparency provided by `USAspending.gov`. First, the site lacks **context**: data isn't cross referenced, so businesses that have changed their names are hard to identify and locate uniquely, and what actual services they provide is even less clear. Context in naming is just one aspect of the problem: finding the origin of contracts, such as those in TX-22, is impossible today. We can't draw any conclusions from these localized spending increases without data on the bid process. In addition, the site contains numerous expenditures that look suspicious at first glance, but may be benign in reality. Changes in spending priorities can't be cross-referenced with with policy changes using the data that's available today. In many instances, the trend in spending visible in the database raises more questions than it answers, making the job of separating the wheat from the chaff harder still.

The second serious issue is that the site provides what I call "transparency as a service"—it provides an interface to *explore* the data, not retrieve it. Performance is one aspect: an uncached search via Google takes about 250ms whereas a similar search on the much smaller database at `USAspending.gov` can take tens of seconds, even minutes. This isn't a fundamental limitation, but it highlights a poor appropriation of re-

[1] The details I present in each example are the findings of an ad-hoc but detailed study I conducted, both manually and using automated routines, between 12/07 and 03/08. The numbers are directly from `USAspending.gov`.

[2] Tracking down these subsidiaries is of interest to the public because photographs from private airstrips indicate that these businesses may have been involved in rendition; the airplanes themselves were operated under possibly dozens of different business names, some of which were not registered within the United States.

sources. Google and other search engines are designed to respond quickly to search requests across vast data sets. It's wasteful to reproduce this functionality. Indeed, while `USAspending.gov` itself was "created" by *S. 2590*, the main value of that legislation was to make data available; the software that powers the service was written originally for OMB Watch's `FedSpending.org`. Furthermore, access to the data is limited to a narrow API and search functionality.

There are initial steps we can take to fix these two specific issues. First, to address unidentifiability of entities in the database, we can assign each entity a unique identifier that is fixed across all federal transparency databases. This identifier enables lookup of, for example, a business that appears in an entry at `USAspending.gov`. (This assumes that we also publish the list of all businesses registered to receive government contracts.) Similarly, individuals who fill out paperwork to register businesses would be assigned unique identifiers. To remedy the lack of context for potentially earmark-driven contracts, we can tag each earmark with information about its author and assign the earmark itself a unique identifier. By tagging any contracts derived from this earmark with the earmark's unique identifier, citizens can determine the origin of porkbarrel contracts. While these are only examples of how we might provide more context in the current system, they are important first steps to improving the usefulness of the data that is available today.

Second, we must recognize that the **data** itself is the key—the challenge of transparency is that data is unavailable to the public, or, when it is available, it's not in a format that can be easily and quickly searched, manipulated, and duplicated. While there are several natural analyses one might perform on a fully integrated and cross-referenced dataset, such as the computation of an approximation of vertex cover, clique, and subgraph isomorphism on the graph of entities [4], our job is not to perform such analyses, but rather to enable them. The data itself, in a persistent, machine parseable, and open form is more important than the service that surrounds it. If we made `USAspending.gov` crawlable, we would enable search engines such as Google to perform queries across its data, providing a front-end and simultaneously obviating `USAspending.gov`'s own fledgling search interface. However, we would need to take this one step further. Since modern search engines trade off flexibility of search with performance, as do their back-end designs [1], they will remain insufficient for a small but important subset of users who wish to delve deeper into the data and perform complex, relational queries. To serve these users, we must publish the data itself in an archive format and distribute it using tools such as BitTorrent or CoBlitz. Furthermore, once users download snapshots of data sets, they need to be able to receive updates as they come in.[3]

Barack Obama has put forward a serious vision of technology-based transparency to help citizens observe the inner workings of their government. Lessig and Trippi have also recently explored this area, and describe the goal succinctly: to "end corruption in America's congress" (and government, more broadly) [2]; one of the main weapons in their arsenal is transparency. However, there are several necessary but *insufficient* conditions for transparency to have the stated, desired effect: a) there are many eyes watching; b) everyone knows that there are many eyes watching; c) some eyes look deeply into the data; d) the data remains up to date; e) the data is presented with full context; and f) the data can be trusted to be authentic.

Transparency itself is not enough. Ideally, it induces members of the government to act on the public's behalf, but transparency can only affect elected officials or moneyed interests if there are consequences to public misbehavior. Therefore, systems of transparency must move beyond a curiosity and provide, to coopt a phrase, "actionable intelligence" that can be used to hold people accountable. To this end, the data itself needs both greater depth and wide dissemination. Government must embrace the needs of those who actually use the data and solicit their input when designing and refining transparency systems.

Finally, the systems themselves need to be broader in scope if they are to provide the necessary context to be useful. For example, we might take a cue from Michael Pollan [3] and reframe the goal of federal funding transparency to aim to reveal the life of a dollar, from birth to death, as it winds its way through government; the journey here is far more important than the destination, as it is more important to expose how, why, when, and by whom each dollar is sent to its final fate.

[3]The emphasis here is to provide the data quickly and easily—using peer to peer file distribution—rather than provide a *service* for high-bandwidth data distribution. Since the data sets in question change rapidly over time, referencing a snapshot in time is crucial to data analysis. Therefore, all data that is available must be referenceable via permanent links to each snapshot.

References

[1] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A Distributed Storage System for Structured Data. In *Proceedings of USENIX/ACM OSDI*, 2006.

[2] L. Lessig and J. Trippi. Change Congress. `http://change-congress.org`.

[3] M. Pollan. *The Omnivore's Dilemma: A Natural History of Four Meals*. Penguin Press, 2006.

[4] V. V. Vazirani. *Approximation Algorithms*. Springer, 2004.