February 3, 2010

By e-mail to
Vivek Kundra, Federal Chief Information Officer
Eisenhower Executive Office Building
Room 269
1650 Pennsylvania Avenue, NW
Washington, DC 20502

Dear Mr. Kundra,

As advocates for government openness, we support the Administration's efforts to provide the public with access to information through Data.gov . We are eager to work with you to ensure the success of Data.gov and, in that spirit, write to raise our concerns with the datasets submitted by agencies to fulfill their requirement under the Open Government Directive to post three high value datasets by January 22, and to offer constructive suggestions for improving their usefulness.

As an overall recommendation, we urge you to add public representatives to the Open Government Initiative interagency working committee and ask the committee to address the problems and recommendations identified below.

Release Format and Usability by the Public

We understand one of the primary purposes of Data.gov is to enable the technology community and transparency advocates to most effectively use the data to make a direct impact on the daily lives of the American people. The format of the data plays a key role in its usability; many within the community of advocates who re-use and repackage government data would prefer data in CSV format, rather than the XML format in which many of the posted databases are provided. Accordingly, we recommend that you strike an appropriate balance between formats (such as XML) that serve the coding community and web-based presentations by agencies that can be used and understood by the general public.

In addition, some of the currently posted files are quite large, ranging upward to several hundred megabytes. Their large size undermines their usefulness for most people or organizations. The large number of currently posted datasets also makes it difficult to find a particular database of interest. We therefore recommend that if a Data.gov dataset is available from an agency through a web-based interface, Data.gov link to that interface on the dataset's Data.gov landing page. For a consumer looking for information on a car seat, for example, it would be far easier to search the Department of Transportation's online database rather than scrolling through screen after screen of raw data in XML format. Additionally, as agencies continue to post datasets to Data.gov, efforts should be made to identify those of greatest public interest that lack such interfaces and develop web interfaces that allow the data to be explored online.

Further, while we agree there is value in aggregating government data in a single site, it is questionable how much the collocation of the currently posted information on Data.gov actually benefits the public. The site is not searchable by topic and does not provide any way to bring together data from different sources on similar topics.

As an enhancement to the organization of the site, we recommend that you use tagging or metadata to enable the public to bring together information on a topic. The thesaurus that USA.gov uses provides a useful example of the needed vocabulary.

Value of Data

The release of the datasets also has prompted discussions about the value and the quality of the released data, and the additional value provided by access to existing data in a new format. We believe repackaging old information is of marginal value, yet that is what many agencies have done with their recent postings on Data.gov. According to the Sunlight Foundation, of 58 datasets posted by major agencies, only 16 were previously unavailable in some format online. This leaves the impression that agencies posted easily available data, the proverbial low-hanging fruit, rather than seriously considering which of their datasets truly are of high value. While these initial postings can be considered a test run, more attention needs to be directed toward ensuring the overall quality and usefulness of the data. In addition, sustained attention should be paid to the possibility of making some of the datasets available as feeds that are constantly up to date, rather than as static datasets that are pulled down and then reposted on an occasional basis.

We recommend that agencies be required to explain why the data is high value by having them designate which of the "high value criteria" the data meets: information that can be used to increase agency accountability and responsiveness; improve public knowledge of the agency and its operations; further the core mission of the agency; create economic opportunity; or respond to need and demand as identified through public consultation. Similarly, we recommend requiring agencies to indicate whether a high value dataset was previously unavailable, available only with a FOIA request, available only for purchase, or available, but in a less user-friendly format. Going forward, this will make it much easier to track how agencies are complying with the other requirements of the Open Government Directive.

While we appreciate the value of data that furthers the mission of an agency, we believe it is equally important to make available to the public data that holds an agency accountable for its policy and spending decisions. We hope to see more datasets of this type available in the near future.

Quality

As is to be expected in efforts of this type, there were a number of glitches--datasets that could not be downloaded or, once downloaded, could not be opened (the Central Contractor Registration FOIA extract from the General Services Administration seems to have caused several users problems). Additionally, some datasets were incomplete (the Hazard Grant Mitigation Program data released by FEMA is missing 23 years of data between 1966 and 1989). Even more troubling, some did not have header rows, and for those that did, their Data.gov pages did not always link to code sheets explaining what those header rows meant. Without this information, the data cannot be used.

We therefore urge the implementation of a responsive feedback mechanism that allows the public to alert an agency that a specific dataset is not working, lacks information, or is missing explanatory material and provides a response to the concerns within a specified time. One way to address this may

be to include an agency contact with the ability to resolve any database problems or provide information about the database. The interagency working group could sample the quality of these agency-specific dialogues to ensure that they are having an impact and to develop recommendations on best practices to improve the responsiveness. Additionally, we strongly recommend that all datasets on Data.gov be directly associated with their code sheets.

Finally, we are concerned with the current lack of public notice when data is removed from the site. We respectfully urge you to note all raw tools and data that are removed from Data.gov, and to provide an explanation for their removal.

Many of the concerns outlined above apply across all or many of the agencies' datasets. Accordingly, we think that standards for handling these types of problems can easily be addressed through the interagency working group and then disseminated amongst the agencies.

We would appreciate the opportunity to discuss this matter in greater detail. Please contact Patrice McDermott, OpenTheGovernment.org, (202-994-332-6736 /pmcdermott@openthegovernment.org) to coordinate a discussion with the signatories to this letter.

Thank you for taking the time to consider our concerns and suggestions. We hope the spirit behind the movement towards transparency — that a government that is open, accountable, and communicative will ultimately be more effective — does not get lost amid the zeal for technology. The White House and its agencies deserve credit for taking this step in the right direction, but more work is needed.

Sincerely,

Gary Bass
Executive Director, OMB Watch

Danielle Brian
Executive Director, Project On Government Oversight (POGO)

Meredith Fuchs
General Counsel, National Security Archive

Ari Schwartz
Vice President and Chief Operating Officer, Center for Democracy and Technology (CDT)

Patrice McDermott
Director, OpenTheGovernment.org

Ellen Miller
Co-founder and Executive Director, Sunlight Foundation

Anne Weismann
Chief Counsel, Citizens for Responsibility and Ethics in Washington (CREW)

cc: Aneesh Chopra
Federal Chief Technology Officer of the United States (CTO)

Norm Eisen
Special Assistant to the President and Special Counsel to the President

Dr. Beth Noveck
Office of Science and Technology Policy, Executive Office of the President

Cass Sunstein
Administrator, Office of Information and Regulatory Affairs, Office of Management and Budget