Derek Willis
Web Developer, The New York Times
Media Access to Government Information Conference

Comments on Question 2: What are the common technical challenges journalists face in making sense of government documents and analyzing government actions, and how could those be overcome?

I very much doubt that the conference organizers intended this, but the fact that our responses to these questions were requested in either PDF or MS Word formats is an excellent example of one of the technical challenges for journalists when dealing with government documents. So in the spirit of openness, I wrote this in Google Docs.

Both journalists and government employees who create and manage information need to know about more than the usual options for the collection and dissemination of information. Part of the technical failure rightly belongs to journalists -- too often, we don't ask or don't know how to ask for information in a way that makes it easy to use. But far too often, government officials are either unaware of their format options or, more perniciously, all too aware and resort to distributing documents in, for example, a locked PDF.

I have been told many times that to release information in a format that would allow it to be copied is not the policy of a government agency. Those government agencies fail to understand what public information is. I also have been the recipient of records that clearly were stored in spreadsheet software but, for purposes of public release, have been printed out, scanned into images and stored as a PDF. Obfuscation and paranoia are not technical challenges, but they contribute to them, forcing journalists to acquire costly software or spend additional time overcoming an artificial roadblock.

This challenge is not due to deficiencies in software produced by any particular company but rather in the understanding of how information can and should be made available to citizens. To the greatest extent possible, government information should be made available in formats that allow its users to copy, sift and reorganize it as they please. In practice, this means favoring text-based and open formats over images, PDFs and closed formats. I am less concerned about how agencies store their data, as long as they are able to export it in common formats or reliable workarounds exist, but there are exceptions to this.

Map data, for example, is commonly stored at the government level in ESRI's shapefile format, which, from the point of view of a journalist, has advantages and disadvantages. ESRI is a large, well-known company with products in use at most government agencies that have geographic data, so it makes sense that GIS data would be provided in that format. But not all newsrooms, and certainly not all journalists, are able to obtain ESRI's software or have access to someone who can easily convert from shapefiles to other formats, such as the KML standard now owned by Google. Some government agencies already produce useful geographic data in KML format, but many others could join that list.

Fixing this situation will require education of journalists and government employees of the benefits and ease of working with open formats. The benefits for journalists are apparent: faster access to information that they can immediately put to use. Training journalists is a time-consuming and inefficient process, but journalists must break out of the mindset that government information only comes in documents.

Doing so means that journalists need to become as comfortable interviewing data as they are interviewing people. The benefits for government may need some more time to explain, but they exist. The Federal Election Commission is a case in point.

Thanks to a commitment to maintaining stable, available and well-documented data, the FEC makes it possible for its users to obtain and analyze information when they want to, even on a late-night deadline. This isn't new-fangled technology; the FEC's FTP site has been operating for years. But the agency operates as if it trusts its data users, not from a defensive standpoint. As a result, the FEC is rightfully seen as an agency that makes it possible for reporters to do their jobs, not as an impediment to that goal.

A more recent, but increasingly significant, technical challenge is that too many government agencies fail to make better use of the best information distribution platform they have: the Internet. In a digital age, some agencies continue to treat all records either as documents, or when they do make data available, it is done as a single dump. In many cases, journalists do not need an entire dataset; they are more likely to want to answer a single question or small set of questions. In those cases where government agencies make this possible, it is usually through a Web form of their own design - one which often is tailored to heavy users such as the regulated community.

Providing Application Programming Interfaces (APIs) to government data via XML or JSON feeds would make it possible for journalists and Web developers to take advantage of government data without having to download and process enormous files. And while adding an API will incur an up-front cost, it will also save agencies employee time handling requests that could be done computer-to-computer.

Yet such APIs are very rare in government, even though they would make it easier for users and journalists to combine disparate data, and would make it possible to build more useful applications from government data. We know this to be true, because in the absence of any meaningful government approach to disseminating legislative data, several outside organizations, including my own, have developed APIs to help spur the use and spread of congressional data.

But in order to do this, we have had to essentially reverse-engineer the Thomas site operated by the Library of Congress, writing fragile HTML parsers that can break should the LoC change the structure of individual pages. So, in order to answer anything beyond the most basic question on legislative matters, a journalist must either spend hours looking up information one

page at a time or be able to write a computer program to parse those pages. There has to be a better way.